

ET36/2005 – Et. Finală/2006 – Sinteza Lucrării
Ministerul Educației și Cercetării
Universitatea Tehnică din Cluj-Napoca
Facultatea de Știința și Ingineria Materialelor
Catedra de Chimie

Programul:	Cercetare de Excelență
Modul: II	Proiecte de Dezvoltare a Resurselor Umane pentru Cercetare
Tipul proiectului:	Proiecte de cercetare de excelență pentru tinerii cercetători
Cod proiect:	ET36/2005
Denumirea proiectului:	Investigații Structurale Integrate pe Compuși Biologic Activi
Etapă:	Finală/2006

Lucrare în extenso

Cuprins

<i>Etape, obiective și activități</i>	<i>– pag. 02</i>
<i>Activități și rezultate</i>	<i>– pag. 03</i>
<i>Publicații</i>	<i>– pag. 09</i>
<i>Potența anti-HIV a derivaților HEPTA și TIBO</i>	<i>– pag. 11</i>
<i>Rezultate ale MDF în SAR</i>	<i>– pag. 17</i>
<i>Coefficientul de partiție octanol-apă pentru fenoli substituiți</i>	<i>– pag. 22</i>
<i>Activitatea erbicidă a triazinelor</i>	<i>– pag. 31</i>
<i>Sistem online pentru SAR cu MDF</i>	<i>– pag. 41</i>
<i>Activitatea antimalarială a chinazolinelor</i>	<i>– pag. 51</i>
<i>Analiza de corelație cu Pearson, Kendall și Spearman</i>	<i>– pag. 62</i>
<i>Coefficientul de partiție octanol-apă pentru bifenili policlorurați</i>	<i>– pag. 82</i>
<i>Concluzii</i>	<i>– pag. 96</i>

Etape, obiective și activități

Pentru etapele anului 2006 obiectivele planificate au fost:

- *Elaborarea modelului de implementare soft și interfață online (Etapa 1);*
- *Culegere de date (Etapa 2);*
- *Elaborarea de modele (Etapa 3).*

Activitățile prevăzute a se desfășura au fost:

- *Achiziție, instalare, testare și configurare aparatura suport (Etapa 1);*
- *Efectuarea de experimente cantitative (Etapa 2);*
- *Măsurări, achiziție și managementul datelor (Etapa 2);*
- *Elaborare metodologie modele (Etapa 3);*
- *Validare modele (Etapa 3);*
- *Participări la manifestări științifice și dobândirea de competențe complementare (Etapetele 1, 2 și 3).*

Activități și rezultate

- Achiziție, instalare, testare și configurare aparatura suport (Etapa 1):**

A fost instalat, testat și configurat serverul web cu adresa IP 193.226.7.140:

193.226.7.140

Up Time

3:10PM up 6 days, 21:07, 0 users, load averages: 0.22, 0.12, 0.04

System Information

Copyright (c) 1992-2005 The FreeBSD Project.

Copyright (c) 1979, 1980, 1983, 1986, 1988, 1989, 1991, 1992, 1993, 1994

The Regents of the University of California. All rights reserved.

FreeBSD 5.4-PRERELEASE #0: Thu Apr 7 13:49:34 EEST 2006

root@j.academicdirect.ro:/usr/src/sys/i386/compile/J

Timecounter "i8254" frequency 1193182 Hz quality 0

CPU	user	nice	system	interrupt	idle
Pentium/P55C (166.19-MHz 586-class CPU)	8.9%	0.0%	7.8%	0.8%	82.6%

- Efectuarea de experimente cantitative; Măsurări, achiziție și managementul datelor (Etapa 2):**

Au fost făcute experimente cantitative, s-au efectuat măsurători și s-a realizat managementul datelor. Rezultatul este prezentat sintetic în continuare, preluat din sistemul de management al datelor realizat:

S-a creat o bază de date conținând câte un table de date pentru fiecare set de molecule.

Iată rezultatul interogării bazei de date SARs:

SQL result

Host: localhost

Database : SARs

Generation Time: Oct 28, 2006 at 02:24

PM

Generated by: phpMyAdmin 2.5.6-rc2 /

MySQL 5.0.22-log

SQL-query: SHOW TABLES LIKE
'%_data';

Rows: 49

Tables_in_SARs (%_data)
19654_data

ET36/2005 – Et. Finală/2006 – Sinteza Lucrării

22583_data
23158c_data
23159e_data
26449t_data
31572_data
3300_data
33504_data
34121bad_data
34121nopt_data
36638_data
41521_data
52344_data
52730_data
DHFR_data
DevMTOp00_data
DevMTOp01_data
DevMTOp02_data
DevMTOp03_data
DevMTOp04_data
DevMTOp05_data
DevMTOp06_data
DevMTOp07_data
DevMTOp08_data
DevMTOp09_data
DevMTOp10_data
DevMTOp11_data
DevMTOp12_data
DevMTOp14_data
DevMTOp15_data
DevMTOp16_data
DevMTOp17_data
DevMTOp18_data
DevMTOp19_data
DevMTOp20_data
DevMTOp21_data
DevMTOp22_data
DevMTOp23_data
DevMTOp24_data
DevMTOp25_data
lChr10_data
MR10_data
PCB_rrf_data
RRC433_lbr_data
RRC433_lkow_data
RRC433_pka_data
Ta395_data

Tox395_data
a_acids_data

ET36/2005 – Et. Finală/2006 – Lucrare in extenso

• **Elaborare metodologie modele (Etapa 3):**

S-a elaborat medodologia de calcul pentru modelele structură-activitate. S-a implementat o interfață care permite evaluarea fiecărui model individual. Imaginea următoare este din interfața elaborată:

MDF Demo Calculator

Molecule filename: 03_mr1003.hin		Distance operator: Topological distance, t Geometrical distance, g	
Atomic property: Cardinality, C Count of directly bounded hidrogen's, H Relative atomic mass, M Atomic electronegativity, E Group electronegativity, G Partial charge, Q			
Interaction model: Rare model and resultant relative to fragment's head, R Rare model and resultant relative to conventional origin, r Medium model and resultant relative to fragment's head, M Medium model and resultant relative to conventional origin, m Dense model and resultant relative to fragment's head, D Dense model and resultant relative to conventional origin, d			
Fragmentation criteria: Minimal fragments, m Maximal fragments, M Szedged distance based fragments, D Cluj path based fragments, P		Linearization operator: Identity (no change), I Inversed I, i Absolute I, A Inversed A, a Logarithm of A, L Logarithm of I, l	
Molecular overall superposing formula: Cond., smallest, m Cond., highest, M Cond., smallest absolute, n Cond., highest absolute, N Avg., sum, S Avg., average, A Avg., S/count(fragments), a Avg., Avg.(Avg./atom)/count(atoms), B Avg., S/count(bonds), b Geom., product, P Geom., mean, G Geom., P ¹ /count(fragments), g Geom., Geom.(Geom./atom)/count(atoms), F Geom., P ¹ /count(bonds), f Harm., sum, s Harm., mean, H Harm., s/count(fragments), h Harm., Harm.(Harm./atom)/count(atoms), l Harm., s/count(bonds), i			
Descriptor (of interaction) formula:			

ET36/2005 – Et. Finală/2006 – Lucrare in extenso

Distance, `D` = d
Inverted distance, `d` = 1/d
First atom's property, `O` = p1
Inverted O, `o` = 1/p1
Product of atomic properties, `P` = p1p2
Inverted P, `p` = 1/p1p2
Squared P, `Q` = p1p2 ^{1/2}
Inverted Q, `q` = 1/p1p2 ^{1/2}
First atom's Property multiplied by distance, `J` = p1d
Inverted J, `j` = 1/p1d
Product of atomic properties and distance, `K` = p1p2d
Inverted K, `k` = 1/p1p2d
Product of distance and squared atomic properties, `L` = d(p1p2) ^{1/2}
Inverted L, `l` = 1/p1p2d
First atom's property potential, `V` = p1/d
First atom's property field, `E` = p1/d ²
First atom's property work, `W` = p1 ² /d
Properties work, `w` = p1p2/d
First atom's property force, `F` = p1 ² /d ²
Properties force, `f` = p1p2/d ²
First atom's property weak nuclear force, `S` = p1 ² /d ³
Properties weak nuclear force, `s` = p1p2/d ³
First atom's property strong nuclear force, `T` = p1 ² /d ⁴
Properties strong nuclear force, `t` = p1p2/d ⁴

Submit Query

• **Validare modele (Etapa 3):**

S-a elaborat o interfață ce permite validarea individuală a rezultatelor fiecărui model obținut. Rezultatele sunt încărcate dintr-un fișier text. Setul se împarte în 2 subseturi: setul Școală și setul Test. Rezultatul rulării aplicației realizate pentru un set de 10 molecule la care s-a măsurat refracția molară MR este redat mai jos:

Set file: MR10.txt
 Training set count: 6
 Training set: mr07 mr01 mr04 mr08 mr06 mr09
 Test set: mr02 mr03 mr05 mr10
 Training set data:

Mol	IGDmSMt	lAmrfEt	Y
mr07	4.0972	1.0740	43.005
mr01	4.2161	1.1978	35.808
mr04	4.4888	1.3004	34.911
mr08	3.8446	8.8195e-1	52.029
mr06	4.4222	1.3185	31.636
mr09	3.9102	9.2678e-1	49.971

QSAR/QSPR: Y_EST = 17.890+28.126*IGDmSMt+-83.972*lAmrfEt
 Coefficient of determination r² = 0.99992008084819 (n = 6)
 Fisher test value F = 18693
 Probability of wrong (from F) p_F = 0.00007 % (7.1870589724021E-07)
 Test set data:

Mol	IGDmSMt	lAmrfEt	Y
mr02	4.0257	1.0788	40.524

ET36/2005 – Et. Finală/2006 – Lucrare in extenso

mr03	4.7052	1.4330	30.03
mr05	4.3624	1.3271	29.222
mr10	3.7941	7.8878e-1	58.323

Coefficient of determination $r^2 = 0.99999289485894$ (n = 4)

Fisher test value F = 10785

Probability of wrong (from F) $p_F = 0.68084 \%$ (0.0068084118091624)

• **Participări la manifestări științifice și dobândirea de competențe complementare (Etapele 1, 2 și 3):**

S-a participat cu lucrări științifice la următoarele conferințe:

1. Sorana Daniela BOLBOACĂ, Ștefan ȚIGAN, Lorentz JÄNTSCHI, *Molecular Descriptors Family on Structure-Activity Relationships on anti-HIV-1 Potencies of HEPTA and TIBO Derivatives*, **European Federation for Medical Informatics Special Topic Conference**, April 6-8, Conference Proceedings Integrating Biomedical Information: From eCell to ePatient (ISBN 3-89838-0722-6 :Aka, ISBN 1-58603-614-9 :IOS Press, ISBN 973-625-303-1 :Ed. Politehnica Timișoara), p. 110-114, **Romania, 2006**, Timișoara. http://lori.academicdirect.org/conferences/Timisoara_2006.pdf
2. Lorentz JÄNTSCHI, Sorana Daniela BOLBOACĂ, *Molecular Descriptors Family on Structure-Activity and Structure-Property Relationships: Results*, **SizeMat: Workshop on Size-Dependent Effects in Materials for Environmental Protection and Energy Application, Specific Support Action, FP6: EC-INCO-CT-2005-016414**, May 25-27, Workshop Proceedings, p. 14-15, **Bulgaria, 2006**, Varna. http://lori.academicdirect.org/conferences/SizeMat_Abstracts.pdf
http://lori.academicdirect.org/conferences/SizeMat_AO4_JantschiL.pdf
3. Lorentz JÄNTSCHI, Sorana-Daniela BOLBOACĂ, *Modeling the Octanol-Water Partition Coefficient of Substituted Phenols by the Use of Structure Information*, **Third Humboldt Conference on Computational Chemistry**, June 24-28, Conference Proceedings, ISBN 954-323-199-0 then 978-954-323-199-7, p. 65, **Bulgaria, 2006**, Varna. http://lori.academicdirect.org/conferences/3HCCC_Varna_June_2006_1.pdf
4. Ștefan ȚIGAN, Lorentz JÄNTSCHI, Sorana-Daniela BOLBOACĂ, *Modeling Herbicidal Activity of a Substituted Triazines Class by Integration of Compounds Complex Structural Information*, **XXIII International Biometric Conference**, July 16-21, e-Proceedings, **Canada, 2006**, Montreal. http://lori.academicdirect.org/conferences/IBC_06.pdf

ET36/2005 – Et. Finală/2006 – Lucrare in extenso

5. Lorentz JÄNTSCHI, Sorana-Daniela BOLBOACĂ, *Online System for Molecular Descriptors Family on Structure-Activity Relationships: Assessment and Characterization of Biologic Active Compounds*, **6th European Conference on Computational Chemistry**, September 3-7, Book of abstract, **Slovakia, 2006**, Bratislava.
http://lori.academicdirect.org/conferences/6ECCC_OnlineSystem.pdf
6. Lorentz JÄNTSCHI, Sorana BOLBOACĂ, *Modelling the Inhibitory Activity on Carbonic Anhydrase IV of Substituted Thiadiazole- and Thiadiazoline- Disulfonamides: Integration of Structure Information*, *New Frontiers in Medicinal Chemistry*, **1st European Chemistry Congress**, 2006 August 27-31, Budapest, Hungary.
http://lori.academicdirect.org/conferences/EuChemC2006_1_list.pdf
http://lori.academicdirect.org/conferences/EuChemC2006_1.pdf
7. Lorentz JÄNTSCHI, Sorana BOLBOACĂ, *Pearson versus Spearman, Kendall's Tau Correlation Analysis on Structure-Activity Relationships of Biologic Active Compounds*, **ISCB27 - International Society for Clinical Biostatistics**, 2006 August 27-31, Geneva, Switzerland, Conference Program, abstract no. 274.
<http://lori.academicdirect.org/conferences/Geneva06.pdf>
http://lori.academicdirect.org/conferences/ISCB_abs274-Jantschi.pdf
8. Lorentz JÄNTSCHI, Sorana-Daniela BOLBOACĂ, *Modelling the Inhibitory Activity on Carbonic Anhydrase I of Some Substituted Thiadiazole- and Thiadiazoline- Disulfonamides: Integration of Structure Information*, **17th European Symposium on Computer Aided Process Engineering**, accepted, 2007 May 27-30, Bucharest, Romania.
http://lori.academicdirect.org/conferences/Escape17_AbsList.pdf
http://lori.academicdirect.org/conferences/Escape17_Paper.pdf

Publicații

1. Lorentz JÄNTSCHI, Sorana BOLBOACĂ, *Molecular Descriptors Family on Structure Activity Relationships 5. Antimalarial Activity of 2,4-Diamino-6-Quinazoline Sulfonamide Derivates*, **Leonardo Journal of Sciences**, AcademicDirect, Internet, [Issue 8](#), 77-88, 2006, ISSN 1583-0233, [DOAJ](#), http://ljs.academicdirect.ro/A08/77_88.pdf
2. Lorentz JÄNTSCHI, Sorana BOLBOACĂ, *Molecular Descriptors Family on Structure Activity Relationships 6. Octanol-Water Partition Coefficient of Polychlorinated Biphenyls*, **Leonardo Electronic Journal of Practices and Technologies**, AcademicDirect, Internet, [Issue 8](#), 71-86, 2006, ISSN 1583-1078, [DOAJ](#), http://lejpt.academicdirect.ro/A08/71_86.pdf
3. Sorana BOLBOACĂ, Claudia FILIP, Ștefan ȚIGAN, Lorentz JÄNTSCHI, *Antioxidant Efficacy of 3-Indolyl Derivates by Complex Information Integration*, **Clujul Medical**, Editura Iuliu Hatieganu, Cluj-Napoca, Issue LXXIX(2), 204-209, 2006, ISSN 1222-2119, [N.U.R.C. Class B](#), http://lori.academicdirect.org/articles/ClujulMedical06_1.pdf
4. Sorana BOLBOACĂ, Lorentz JÄNTSCHI, *Pearson Versus Spearman, Kendall's Tau Correlation Analysis on Structure-Activity Relationships of Biologic Active Compounds*, **Leonardo Journal of Sciences**, AcademicDirect, Internet, [Issue 9](#), 179-200, 2006, ISSN 1583-0233, [DOAJ](#), http://ljs.academicdirect.ro/A09/179_200.pdf
5. Sorana BOLBOACĂ, Lorentz JÄNTSCHI, *Molecular Descriptors Family on Structure-Activity Relationships: Modeling Herbicidal Activity of Substituted Triazines Class*, **Bulletin of University of Agricultural Sciences and Veterinary Medicine - Agriculture**, AcademicPres, Romania, Volume 62, 35-40, 2006, ISSN 1454-2382, [N.U.R.C. Class B](#), [http://lori.academicdirect.org/articles/BUSAMV_06\(62\).pdf](http://lori.academicdirect.org/articles/BUSAMV_06(62).pdf)
6. Lorentz JÄNTSCHI, Sorana-Daniela BOLBOACĂ, *Modeling the Octanol-Water Partition Coefficient of Substituted Phenols by the Use of Structure Information*, **International Journal of Quantum Chemistry**, trimisă spre publicare, http://lori.academicdirect.org/articles/IJQC_Jantschi&Bolboaca.pdf
7. Lorentz JÄNTSCHI, Sorana-Daniela BOLBOACĂ, *Online System for Molecular Descriptors Family on Structure-Activity Relationships: Assessment and Characterization of Biological Active Compounds*, **International Journal of Quantum Chemistry**, trimisă spre publicare, http://lori.academicdirect.org/articles/IJQC_6ECCC_OnlineSystemMDFSAR.pdf

ET36/2005 – Et. Finală/2006 – Lucrare in extenso

8. Lorentz JĂNTSCHI, Sorana Daniela BOLBOACĂ, *Counting Polynomials on Regular Structures*, **MATCH**, trimisă spre publicare,
http://lori.academicdirect.org/articles/MATCH_CountingPolynomials.pdf
9. Lorentz JĂNTSCHI, Violeta POPESCU, *Relatii structura-activitate în prezicerea toxicitatii asupra Tetrahymena Pyriformis a substituentilor în pozitia para pe fenol*, **Revista de Chimie**, trimisă spre publicare,
http://lori.academicdirect.org/articles/Paper_RRC_433_RevChimie_.pdf
10. Horea Iustin NAȘCU, Lorentz JĂNTSCHI, *Chimie analitică și instrumentală (in Romanian)*, AcademicDirect & AcademicPres, **Internet & Cluj-Napoca**, 320 p., ISBN(10) 973-744-046-3 & ISBN(13) 978-973-744-046-4 (AcademicDirect) && ISBN (10)973-86211-4-3 & ISBN(13) 978-973-86211-4-5 (AcademicPres), **2006 (November)**, în curs de apariție. <http://ph.academicdirect.org/>

Molecular Descriptors Family on Structure-Activity Relationships on anti-HIV-1 potencies of HEPT and TIBO derivatives

Abstract

A new developed methodology of Structure-Activity Relationships (SAR) was applied on a set of 57 compounds with known inhibition activity of immunodeficiency virus type 1. The methodology uses an original family of molecular structure descriptors called Molecular Descriptors Family. With a set of multiple linear regression analysis programs, the whole set of MDF members were crossed in order to find the best SAR model. The obtained model allows making important remarks on structure-activity links. The disadvantage of time consuming to analyze the entire set of descriptors is compensated by better structure-activity relationships.

Keywords

HIV-1 inhibitors, Structure Activity Relationships (SAR), Molecular Descriptors Family (MDF)

Introduction

Two different types of human immunodeficiency viruses (HIV-1 and HIV-2) differing in nucleotide and amino-acid sequences are responsible by the acquired immunodeficiency syndrome, but the HIV-1 type is most predominant [1].

A previous study analyzed the HEPT and TIBO derivatives potencies on HIV-1 [2] using quantitative structure-activity relationships methodology. The results obtained by Toporov & all are:

$$\begin{aligned} n &= 57; r = 0.9397; s = 0.520; F = 416 \text{ (all compounds)} \\ n &= 37; r = 0.9426; s = 0.513; F = 279 \text{ (training set)} \\ n &= 20; r = 0.9408; s = 0.547; F = 139 \text{ (test set)} \end{aligned} \quad (1)$$

[1] Gallo RC, Reitz Jr. MS. The first human retroviruses: are there others? Microbiological Sciences 1985; 2(4): 97-104.

[2] Toropov AA, Toropova AP, Nesterova IV, Nabieva O.M. Comparison of QSAR models of anti-HIV-1 potencies based on labeled hydrogen filled graph and graph of atomic orbitals. Journal of Molecular Structure: THEOCHEM. 2003;640(1-3):175-81.

ET36/2005 – Et. Finală/2006 – Lucrare in extenso

where n is size of the sample; r is the correlation coefficient; s is standard error and F is Fisher parameter.

Starting with the integration of complex structure information of HEPTA and TIBO derivatives, the aim of the research was to evaluate the ability of molecular descriptor family structure-activity relationships in modeling of the inhibition effectiveness against HIV-1.

Material and Method

A number of nineteen HEPT derivatives and thirty-eight TIBO derivatives with inhibition properties on HIV-1 were included into the study. The effectiveness in inhibiting HIV-1 of HEPT and TIBO derivatives (two groups of reverse transcriptase inhibitors) was taken from a previous paper [3] and is expressed as the concentration of compound required to achieve 50% protection of MT-4 cells against the virus (called $\log(10^6/C_{50})$).

The use of a new original set of molecular descriptors, called Molecular Descriptors Family (MDF) into a Quantitative Structure-Activity Relationship was applied in order to study the inhibiting HIV-1 activity of 19 HEPT and 38 TIBO compounds. The steps of molecular descriptor family on structure activity relationship (MDF SAD) modeling are [4]:

- Step I: Sketch of HEPT and TIBO compounds by the use of HyperChem software [5];
- Step II: Create the file with measured inhibiting HIV-1 activity (Y_{C50}) of HEPT and TIBO compounds;
- Step III: Generate the MDF members based on topological and geometrical representations of the compounds. There were identified 296965 MDF members with real and not identical values from which only 95277 were distinct each from other. More, considering also the withdrawing of planar dependencies (one descriptor is dependent on other two) it remains only 84408.

[3] Ganzalez OG, Murray JS, Peralta-Inga Z, Politzer P. Computed molecular surface electrostatic potentials of two groups of reverse transcriptase inhibitors: Relationships to anti-HIV-1 activities. *Int. J. Quantum Chem.* 2001;83:115–21

[4] Jäntschi L. Molecular Descriptors Family on Structure Activity Relationships 1. The review of Methodology, *Leonardo Electronic Journal of Practices and Technologies.* AcademicDirect. 2005;6:76-98.

[5] ***, HyperChem, Molecular Modelling System [Internet page]; ©2003, Hypercube, Inc., available at: <http://hyper.com/products/>

ET36/2005 – Et. Finală/2006 – *Lucrare in extenso*

- Step IV: Finding the SAR models for HEPT and TIBO compounds. The selected members enter into multiple linear regression analysis. Mono-varied and multi-varied models were applied. At the end of all pair's computations the best QSAR models were selected and presented here. Note that for bi-varied model, 3562313028 pairs enters into bi-varied regression model and a multiple of enters into tri- and more varied models.
- Step V: Validation of the obtained SAR models were performed through computing the cross-validation leave-one-out correlation score [6], and the difference between this parameter and the squared correlation coefficient.
- Step VI: Analyze the selected SAR model and comparing it with previous reported model.

Results

The best performing SAR (five-varied model) was selected and is presented here. The selection of the best performing five-varied model was made first after the greatest squared correlation coefficient and then after the greatest values of cross-validation leave-one-out (loo) score ($r^2_{cv(loo)}$).

The models have the following equation:

$$\hat{Y} = 17.7 - 7.11 \cdot InMdTHg - 1.23 \cdot IFDMwEt + 8.36 \cdot AiMrKQt + 6.59 \cdot 10^5 \cdot ImDMtQt - 5.98 \cdot IIMdEMg \quad (2)$$

where \hat{Y} is predictor of measured inhibition activity (Y_{C50}) and $InMdTHg$, $IFDMwEt$, $AiMrKQt$, $ImDMtQt$, and $IIMdEMg$ are molecular descriptors.

The characteristics associated with the above-described models are in table 1 and is graphically represented in figure 1.

Table 1. Statistics associated with the five-varied model

Characteristic	Notation	SAR Model (eq. 2)
Correlation coefficient	r	0.9579
Squared correlation coefficient	r^2	0.9175
Adjusted squared correlation coefficient	r^2_{adj}	0.9094
Standard error of estimated	S_{est}	0.4521
Fisher parameter	F_{est}	113
Probability of wrong model	$p_{est}(\%)$	$2.14 \cdot 10^{-24}$
t parameter for intercept	t_{int}	22.33
p-values	$p_{tint}(\%)$	$5.52 \cdot 10^{-26}$
95% CI (confidence interval) [lower 95%; upper 95%]	$95\% CI_{int}$	[16.13, 19.32]
t parameter for InMdTHg descriptor	$t_{InMdTHg}$	-9.27

[6] ***, Leave-one-out Analysis. ©2005, Virtual Library of Free Software, available at:

http://vl.academicdirect.org/molecular_topology/mdf_findings/loo/

Associated p-value	$p_{\text{InMdTHg}} (\%)$	$1.63 \cdot 10^{-10}$
95% CI [lower 95%; upper 95%]	$95\% \text{CI}_{\text{InMdTHg}}$	[-8.65, -5.57]
t parameter for IFDMwEt descriptor	t_{IFDMwEt}	-12.43
Associated p-value	$p_{\text{IFDMwEt}} (\%)$	$4.75 \cdot 10^{-15}$
95% CI [lower 95%; upper 95%]	$95\% \text{CI}_{\text{IFDMwEt}}$	[-1.43, -1.04]
t parameter for AiMrKQt descriptor	t_{AiMrKQt}	9.58
Associated p-value	$p_{\text{AiMrKQt}} (\%)$	$5.43 \cdot 10^{-11}$
95% CI [lower 95%; upper 95%]	$95\% \text{CI}_{\text{AiMrKQt}}$	[6.61, 10.11]
t parameter for ImDMtQt descriptor	t_{ImDMtQt}	6.86
Associated p-value	$p_{\text{ImDMtQt}} (\%)$	$9.22 \cdot 10^{-7}$
95% CI [lower 95%; upper 95%]	$95\% \text{CI}_{\text{ImDMtQt}}$	$[4.66 \cdot 10^5, 8.52 \cdot 10^5]$
t parameter for IIMdEMg descriptor	t_{IIMdEMg}	-7.07
Associated p-value	$p_{\text{IIMdEMg}} (\%)$	$4.15 \cdot 10^{-7}$
95% CI [lower 95%; upper 95%]	$95\% \text{CI}_{\text{IIMdEMg}}$	[-7.68, -4.29]
Cross-validation leave-one-out (loo) score	$r^2_{\text{cv(loo)}}$	0.8997
Fisher parameter for loo analysis	F_{pred}	91
Probability of wrong model for loo analysis	$p_{\text{pred}} (\%)$	$< 10^{-17}$
Standard error for loo analysis	S_{loo}	0.4987
The difference between r^2 and $r^2_{\text{cv(loo)}}$	$r^2 - r^2_{\text{cv(loo)}}$	0.0178

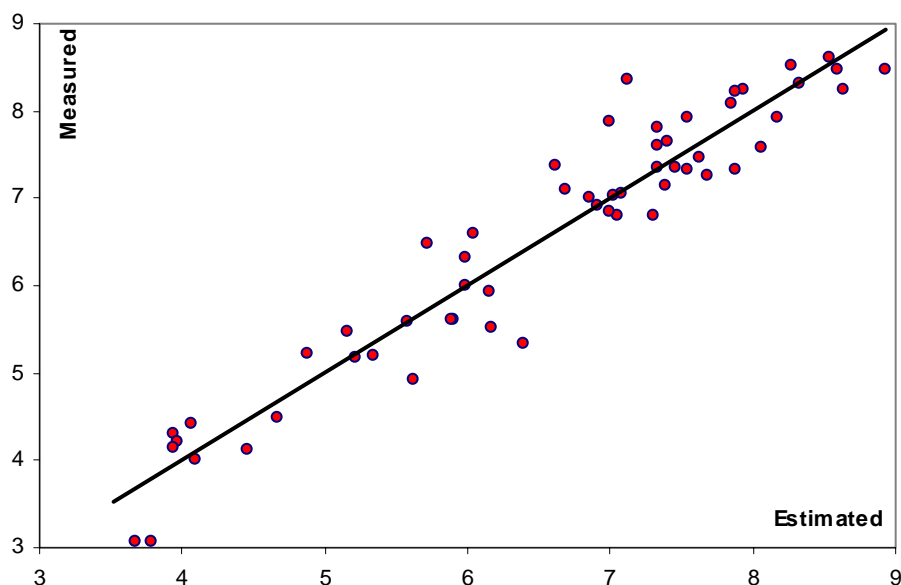


Figure 1. Measured inhibition activity (Measured) vs. estimated (Estimated) with five-varied SAR model

Assessment of the MDF SAR model was performed by applying a correlated correlation analysis, which took into consideration the five-varied SAR models and compared it with the best performing (model with four variables, $r = 0.9397$, $n = 57$ – equation 1) previous reported model [2] by the use of Steiger's Z test. The results of comparison are in table 2.

Table 2. The Steiger's Z test results

Characteristic	Value
$r(Y_{C50}, \hat{Y}_{SAR})$	0.9579
$r(Y_{C50}, \hat{Y}_{Previous})$	0.9399
$r(\hat{Y}_{SAR}, \hat{Y}_{Previous})$	0.9252
Steiger's Z test parameter	1.3462
$p_{Steiger's Z} (\%)$	0.0891

Discussions

The selected best found five-varied SAR model of HEPT and TIBO QSAR HIV-1 inhibiting activity shows that atoms mass and attached hydrogen's has significance on activity behavior using geometrical model of the molecule (InMdTHg and lIMdEMg). Partial charge has significance on activity behavior using strictly topological model. More, two descriptors use the **Qt** association in the selected best found five-varied model (AiMrK**Qt** and ImDMt**Qt**). The atomic and group electronegativity, as a composed property tends with increasing of number of descriptors to be replaced by more accurate properties: attached hydrogen's, partial charge and mass. Thus, if bi-varied model has only atomic and group electronegativity as atomic descriptors, in tri-varied model disappear one electronegativity based descriptor and appear one attached hydrogen's based and one partial charge based, and for five-varied are two descriptors based on partial charge, one based on attached hydrogen's and one based on atomic mass.

Looking at the five-varied model, we can say that the inhibitory activity it is of molecular topology as well as molecular geometry and depend on partial change of molecule, molecular mass and number of bounded hydrogen's. The values of squared correlation coefficient ($r^2 = 0.9175$), the student parameter, associated p-values and 95% confidence intervals (see table 1) demonstrate the goodness of fit of the five-varied MDF SAR model. The power of the five-varied model in prediction of the inhibitory activity of compounds is demonstrate by the cross-validation leave-one-out correlation score ($r_{cv(100)}^2 = 0.8997$). The stability of the best performing five-varied MDF SAR model is give by the difference between the squared correlation coefficient and the cross-validation leave-one-out correlation score ($r^2 - r_{cv(100)}^2 = 0.0178$).

Comparing with previous reported model (equations (1)) [2], our model (equations (2)) is better ($r^2 = 0.918$ – see table 1 and equation 1). At modeling level, our approach is more software independent than previous reported. We use software dependent procedures only for constructing a basic geometrical model of the molecules and compute the partial

ET36/2005 – Et. Finală/2006 – Lucrare in extenso

charge distribution inside the molecules. We do not “optimize” the geometrical shape according to an arbitrary chosen model and/or algorithm.

Starting with the knowledge learned from the studied set, inhibition property of new compound from the same class can be predicted by the use of an original software, which is available at the following address:

http://vl.academicdirect.org/molecular_topology/mdf_findings/sar/

Thus, the software is able to predict the inhibitory activity of new compounds from the same class with low costs.

It can be concluded that the use of MDF for SARs finding on HIV-1 potent compounds offers accurate models and allow making of important remarks about structure-activity links.

Molecular Descriptors Family on Structure-Activity and Structure-Property Relationships: Results

Purpose

To present the results obtained by utilization of an original approach called **Molecular Descriptors Family** (MDF) on **Structure-Property** (SPR) and **Structure-Activity Relationships** (SAR) applied on different classes of chemical compounds and its usefulness as precursors of models elaboration of new chemical compounds with better properties and/or activities.

Methodology

Molecular Descriptors Family

- Preparing chemical compounds for molecular modeling
- Generating the molecular descriptors family
- Finding the MDF SPR/SAR models
- Validating the MDF SPR/SAR models
- Comparing the MDF SPR/SAR models with previous reported model(s)

Materials

Set name	Observed/Measured Property/Activity
IChr	retention chromatography index
PCB_rrf	relative response factor
23159	octanol/water partition coefficients
23159e	octanol/water partition coefficients
PCB_lkow	octanol/water partition coefficient
36638	water activated carbon adsorption
MR10	molar refraction
Ta395	cytotoxicity
52730	toxicity
Tox395	mutagenicity
41521	insecticidal activity
Triazines	herbicidal activity
52344	antioxidant efficacy
26449	antituberculous activity
23151	antimalarial activity
22583	anti-HIV-1 potencies

ET36/2005 – Et. Finală/2006 – Lucrare in extenso

Results: database

vl.academicdirect.org >> localhost >> SARs | phpMyAdmin 2.6.1-pl3 - Microsoft Internet Explorer

Address: http://vl.academicdirect.org/phpMyAdmin/index.php?lang=en-utf-8&server=1&collation_connection=utf8_general_ci

Database:	Table	Rows	Storage Engine	Character Set	Collation	Table Size	Index Size
SARs (194)	MR10_xval	107,692	MyISAM	latin1_swedish_ci		8.3 MB	-
	MR10_yval	107,692	MyISAM	latin1_swedish_ci		4.9 MB	-
	PCB_rrf_data	209	MyISAM	latin1_swedish_ci		2.8 KB	-
	PCB_rrf_tmpx	131,328	MyISAM	latin1_swedish_ci		212.1 MB	-
	PCB_rrf_xval	62,873	MyISAM	latin1_swedish_ci		100.3 MB	-
	PCB_rrf_yval	62,873	MyISAM	latin1_swedish_ci		2.9 MB	-
	RRC433_lbr_data	30	MyISAM	latin1_swedish_ci		1.3 KB	-
	RRC433_lbr_tmpx	131,328	MyISAM	latin1_swedish_ci		32.7 MB	-
	RRC433_lbr_xval	86,409	MyISAM	latin1_swedish_ci		19.9 MB	-
	RRC433_lbr_yval	86,409	MyISAM	latin1_swedish_ci		4.0 MB	-
	RRC433_lkow_data	30	MyISAM	latin1_swedish_ci		1.3 KB	-
	RRC433_lkow_tmpx	131,328	MyISAM	latin1_swedish_ci		32.7 MB	-
	RRC433_lkow_xval	86,388	MyISAM	latin1_swedish_ci		19.9 MB	-
	RRC433_lkow_yval	86,388	MyISAM	latin1_swedish_ci		4.0 MB	-
	RRC433_pka_data	30	MyISAM	latin1_swedish_ci		1.3 KB	-
	RRC433_pka_tmpx	131,328	MyISAM	latin1_swedish_ci		32.7 MB	-
	RRC433_pka_xval	86,373	MyISAM	latin1_swedish_ci		19.9 MB	-
	RRC433_pka_yval	86,373	MyISAM	latin1_swedish_ci		4.0 MB	-
	Ta395_data	15	MyISAM	latin1_swedish_ci		1.1 KB	-
	Ta395_tmpx	131,328	MyISAM	latin1_swedish_ci		17.7 MB	-
	Ta395_xval	102,608	MyISAM	latin1_swedish_ci		11.8 MB	-
	Ta395_yval	102,608	MyISAM	latin1_swedish_ci		4.7 MB	-
	Tox395_data	14	MyISAM	latin1_swedish_ci		1.1 KB	-
	Tox395_tmpx	131,328	MyISAM	latin1_swedish_ci		16.7 MB	-
	Tox395_xval	103,411	MyISAM	latin1_swedish_ci		11.1 MB	-
	Tox395_yval	103,411	MyISAM	latin1_swedish_ci		4.7 MB	-
	a_acids_data	12	MyISAM	latin1_swedish_ci		1.1 KB	-
	a_acids_tmpx	131,328	MyISAM	latin1_swedish_ci		14.7 MB	-
	a_acids_xval	97,007	MyISAM	latin1_swedish_ci		9.0 MB	-
	a_acids_yval	97,007	MyISAM	latin1_swedish_ci		4.4 MB	-
	molecular_topology	86	MyISAM	latin1_swedish_ci		17.1 KB	-
	194 table(s)	Sum				14,305,820	0 Bytes

Results: web interface

http://vl.academicdirect.org/molecular_topology/ - Microsoft Internet Explorer

Address: http://vl.academicdirect.org/molecular_topology/





The screenshot shows a web browser window displaying the VLFS web interface. The page has a logo in the top right corner and a navigation menu on the left side. The menu items are: [Up](#), [cycles](#), [devel](#), [qsar_qspr_s](#), and [tpaths](#). On the right side of the page, there are several links: [cage_versatile](#), [data_mining](#), [mdf_findings](#), and [speed](#).

http://vl.academicdirect.org/?v= - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address http://vl.academicdirect.org/?v=

vl from AcademicDirect: A gateway to free to use software.

VLFS Powered by    

[SysInfo](#)

[admittance_app](#)

[applied_statistics](#)

[general_chemistry](#)

[medical_informatics](#)

[molecular_dynamics](#)

[molecular_topology](#)

[phpMyAdmin](#)

[phpSysInfo](#)

© 2000, Maintained by Lorentz JÄNTSCH

Results: MDF model

Address http://vl.academicdirect.org/molecular_topology/mdf_findings/j_mdf_demo.php

MDF Demo Calculator

Molecule filename: 03_mr1003.hin	Distance operator: Topological distance, t Geometrical distance, g	Atomic property: Cardinality, C Count of directly bounded hydrogen's, H Relative atomic mass, M Atomic electronegativity, E Group electronegativity, G Partial charge, Q
Descriptor (of interaction) formula: Distance, 'D' = d Inverted distance, 'd' = 1/d First atom's property, 'O' = p1 Inverted O, 'o' = 1/p1 Product of atomic properties, 'P' = p1p2 Inverted P, 'p' = 1/p1p2 Squared P, 'Q' = p1p2 ^{1/2} Inverted Q, 'q' = 1/p1p2 ^{1/2} First atom's Property multiplied by distance, 'J' = p1d Inverted J, 'j' = 1/p1d Product of atomic properties and distance, 'K' = p1p2d Inverted K, 'k' = 1/p1p2d Product of distance and squared atomic properties, 'L' = d(p1p2) ^{1/2} Inverted L, 'l' = 1/p1p2d First atom's property potential, 'V' = p1/d First atom's property field, 'E' = p1/d ² First atom's property work, 'W' = p1 ² /d Properties work, 'w' = p1p2/d First atom's property force, 'F' = p1 ² /d ² Properties force, 'f' = p1p2/d ² First atom's property weak nuclear force, 'S' = p1 ² /d ³ Properties weak nuclear force, 's' = p1p2/d ³ First atom's property strong nuclear force, 'T' = p1 ² /d ⁴ Properties strong nuclear force, 't' = p1p2/d ⁴		Interaction model: Rare model and resultant relative to fragment's head, R Rare model and resultant relative to conventional origin, r Medium model and resultant relative to fragment's head, M Medium model and resultant relative to conventional origin, m Dense model and resultant relative to fragment's head, D Dense model and resultant relative to conventional origin, d
Fragmentation criteria: Minimal fragments, m Maximal fragments, M Szeged distance based fragments, D Cluj path based fragments, P	Molecular overall superposing formula: Cond., smallest, m Cond., highest, M Cond., smallest absolute, n Cond., highest absolute, N Avg., sum, S Avg., average, A Avg., S/count(fragments), a Avg., Avg.(Avg./atom)/count(atoms), B Avg., S/count(bonds), b Geom., product, P Geom., mean, G Geom., P ¹ /count(fragments), g Geom., Geom.(Geom./atom)/count(atoms), F Geom., P ¹ /count(bonds), f Harm., sum, s Harm., mean, H Harm., s/count(fragments), h Harm., Harm.(Harm./atom)/count(atoms), I Harm., s/count(bonds), i	Linearization operator: Identity (no change), I Inversed I, l Absolute I, A Inversed A, a Logarithm of A, L Logarithm of I, l

Submit Query

ET36/2005 – Et. Finală/2006 – Lucrare in extenso

Results: comparisons

Set name	Previous reported SAR			MDF SAR			
	r ²	n	v	r ²	r ² _{cv(loo)}	n	v
IChr10	0.900	10	2	0.999	0.999	10	2
PCB_rrf	-	-	-	0.628	0.619	209	1
				0.693	0.682	209	2
				0.737	0.717	209	4
PCB_lkow	-	-	-	0.873	0.87	206	1
				0.89	0.885	206	2
				0.917	0.909	206	4
36638	0.967	16		0.994	0.991	16	3
23159	0.388	18	1	0.755	0.684	18	1
	0.839	18	3	0.982	0.974	18	2
23159e	-	-	-	0.899	0.758	8	1
				0.968	0.898	8	2
Ta395	0.870	13	2	0.977	0.961	15	2
Tox395	0.800	13	2	0.957	0.934	14	2

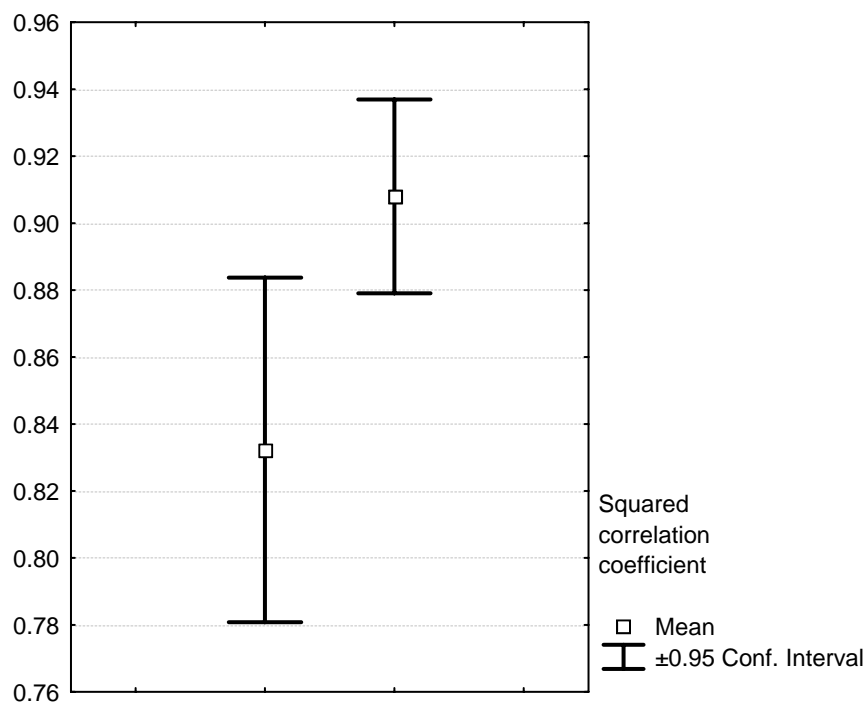
Set name	Previous reported SAR			MDF SAR			
	r ²	n	v	r ²	r ² _{cv(loo)}	n	v
41521	0.913	8	3	0.999	0.998	8	2
	0.985	8	5				
26449	0.991	10	1	0.961	0.954	10	1
	0.998		2	0.99	0.988	10	2
	0.993		4	0.998	0.997	10	4
MR10	0.976	10	2	0.999	0.999	10	2
23151	0.741	16	4	0.997	0.995	16	3
	0.985	13	4				
52344	0.780	8	1	0.904	0.832	8	1
	0.710	8	1	0.999	0.999	8	2
	0.810	8	2	0.999	0.999	8	2
	0.970	8	4	0.999	0.999	8	2

Set name	Previous reported SAR			MDF SAR			
	r ²	n	v	r ²	r ² _{cv(loo)}	n	v
52730	-	-	-	0.966	0.947	10	1
				0.998	0.996	10	2
Triazines	0.970	30	3	0.951	0.946	30	1

ET36/2005 – Et. Finală/2006 – Lucrare in extenso

				0.975	0.971	30	2
				0.983	0.976	30	3
				0.989	0.985	30	4
	0.888	37	5	0.783	0.766	57	2
	0.885	20	5	0.835	0.809	57	3
	0.883	57	5	0.9	0.884	57	4
22583				0.918	0.9	57	5

Results: statistics



Conclusions: SPR

Molecular Descriptors Family on Structure-Property Relationships obtained very good performances in estimation and prediction of compound's properties as are for example retention chromatographic index, molar refraction and water activated carbon organics adsorption.

Conclusions: SAR

Good performances are also obtained by Molecular Descriptors Family on Structure-Activity Relationships in estimation and prediction of compound's activities as are for example toxicity of alkyl metal compounds, insecticidal activity of neonicotinoid compounds, antituberculosic activity of polyhydroxyxanthones, antioxidant efficacy of 3-indolyl derivatives, or antimalarial activity of some 2,4-diamino-6-quinazoline sulfonamide derivatives.

Modeling the Octanol-Water Partition Coefficient of Substituted Phenols by the Use of Structure Information

ABSTRACT

The paper presents the abilities in estimation and prediction of the octanol-water partition coefficient of some para-substituted phenols through the integration of complex structures information by the use of an original molecular descriptors family on the structure-property relationships approach.

The proposed approach uses the complex information obtained from para-substituted phenols structure in order to generate and calculate the molecular descriptors family. The structure-property relationship models were built based on the generated descriptors. The obtained multi-varied models (model with two and four descriptors, respectively) were validated through the assessment of the cross-validation leave-one-out score. The comparison between the multi-varied model with two and four descriptors was performed using Steiger's Z test. The analysis of the statistical characteristics of the obtained models demonstrated that the model with four descriptors has greater abilities to estimate and predict compared with the model with two descriptors. This observation was also sustained by the results of correlated-correlation analysis.

The multi-varied model with four descriptors revealed that the octanol-water partition coefficient of studied para-substituted phenols is likely to be of geometry nature, it is strongly dependent on the partial charges of compounds and group electronegativity and it is in relation with the elastic force.

KEYWORDS

Molecular Descriptors Family on Structure-Property Relationships (MDF-SPR), Octanol-water partition coefficient, Para-substituted phenols

INTRODUCTION

The octanol-water partition coefficient, defined as the ratio of the concentration of a chemical in octanol and in water at equilibrium and at a specified temperature [1] is used by many researchers in quantitative structure-property relationships studies. Partition coefficients are used in medicinal chemistry [2], drug design [3], toxicology [4] and environmental chemistry [5].

ET36/2005 – Et. Finală/2006 – Lucrare in extenso

The literature reported various methods that are able to predict the octanol-water partition coefficient [6] by applying the fragment constant methods [7], by computing van der Waals molecular volume and surface area through analytical and numerical techniques [8], by the use of fuzzy [9], and the neural network approach [10].

An original approach to molecular descriptor family on structure-property relationships (MDF-SPR), method that proved to be able to estimate and predict properties, has been developed [11]. Starting from the successful results obtained by the use of the MDF-SPR methodology on estimation and prediction of retention chromatography index [12], octanol/water partition coefficients [13, 14], water activated carbon adsorption [15], and molar refraction [16], the aim of the research was to study the abilities of the MDF-SPR methodology in estimation and prediction of octanol-water partition coefficient of some para-substituted phenols.

MATERIAL AND METHOD

Para-substituted phenols

A number of thirty para-substituted phenols, previously studied by Schultz T. W. [17] were included in the study.

The generic structure of compounds, their abbreviation (Abb.), the substituent from para position (R), and associated octanol-water partition coefficient, expressed in logarithmic scale are presented in Table 1.

Molecular descriptors family on structure-property relationships methodology

The octanol-water partition coefficient of para-substituted phenols was modeled by the use of the MDF-SPR methodology. The steps followed in the modeling process, described in details in [11], were:

- *3D representation of compounds*: the three-dimensional representations of para-substituted phenols were built up by using **HyperChem** software [18].
- *Creation of measured properties file*: the octanol-water partition coefficient for each para-substituted phenol, expressed in logarithmic scale was stored in *phenols.txt* file.
- *Molecular descriptors family generation and computing*. All thirty compounds were used in the construction and generation of the molecular descriptors family. The algorithm generates the list of molecular descriptors family and associated values for para-substituted phenols, being strictly based on complex information obtained from the compounds structure. In order to discard redundant information, a bias method with a significance

ET36/2005 – Et. Finală/2006 – *Lucrare in extenso*

level equal with 10^{-9} was applied after generation of the molecular descriptors family. Each calculated descriptor has an individual name of seven letters, which express the modality of construction:

- Compound characteristic relative to its geometry (*g*) or topology (*t*) - the 7th letter;
- Atomic property: cardinality (*C*), number of directly bonded hydrogen's (*H*), atomic relative mass (*M*), atomic electronegativity (*E*), group electronegativity (*G*), and the partial charge, semi-empirical Extended Hückel model, Single Point approach (*Q*) – the 6th letter;
- Atomic interaction descriptor – the 5th letter;
- Overlapping interaction model – the 4th letter;
- Fragmentation criterion: the minimal fragments (*m*), the maximal fragments (*M*), the Szeged fragments criterion (*D*), and the Cluj fragments criterion (*P*) [19, 20] – the 3rd letter;
- Cumulative method of fragmentation properties (nineteen functions) – the 2nd letter:
 - Conditional group (four functions): smallest fragmental descriptor value from the array (*m*), highest value (*M*), smallest absolute value (*n*), and highest absolute value (*N*);
 - Average group (five functions): sum of descriptor values (*S*), average mean for valid fragments (*A*), average mean for all fragments (*a*), average mean by atom (*B*), average mean by bond (*b*);
 - Geometric group (five functions): multiplication of descriptor values (*P*), geometric mean for valid fragments (*G*), geometric mean for all fragments (*g*), geometric mean by atom (*F*), and geometric mean by bond (*f*);
 - Harmonic group (five functions): harmonic sum of values (*s*), harmonic mean for valid fragments (*H*), harmonic mean for all fragments (*h*), harmonic mean by atom (*I*), and harmonic mean by bond (*i*);
- Linearization procedure applied in global molecular descriptor generation: identity (*I*), inverse (*i*), absolute (*A*), an inverse of absolute (*a*), natural logarithm of absolute value (*L*), and simple natural logarithm (*l*) - 1st letter.
- *Identification of best performing MDF-SPR models.* The criteria imposed in searching for the best performing models were: the model significance, the values for the correlation and squared correlation coefficients (they were considered performing models if the correlation and/or squared correlation coefficients were closest to +1 or -1), the standard error and the significances of the coefficients.

ET36/2005 – Et. Finală/2006 – Lucrare in extenso

- *Validation of the MDF-SPR models.* The analysis of the predictive abilities of the MDF-SPR models was performed through model validation analysis by computing: the cross-validation leave-one-out (loo) score, the Fisher parameter and its significance for leave-one-out analysis, and the standard error for leave-one-out analysis. In leave-one-out analysis the property of each compound was predicted by the regression equation calculated based on all the other compounds by using the ***Leave-one-out Analysis*** application [21].
- *Analysis of the MDF-SPR models.* The chosen MDF-SPR models were analyzed through computing and interpreting of a number of seven statistical characteristics of the models. The comparison between the multi-varied model with four descriptors and the model with two descriptors was performed through a correlated correlation analysis using Steiger test [22] at a significance level of 5%. The estimation ability of the model with the highest squared correlation coefficient was analyzed in training and test sets using the ***Training vs. Test*** application [23]. Nine situations were analyzed, starting with sample sizes in training sets from fifteen to thirty and corresponding sample sizes in test sets from fifteen to seven.

RESULTS

Two multi-varied MDF-SPR models with two and four descriptors, respectively, proved to have abilities in estimation and prediction of the octanol-water partition coefficient for studied para-substituted phenols. The MDF-SPR models were:

- The MDF-SPR model with two descriptors:

$$\hat{Y}_{2D} = 1.07 + 3.38 \cdot 10^{-3} \cdot isDDkGg - 0.40 \cdot IMmrKQg \quad \text{Eq.(1)}$$

- The MDF-SPR model with four descriptors:

$$\hat{Y}_{4D} = 8.69 \cdot 10^{-2} + 5.56 \cdot 10^{-3} \cdot isDDkGg - 4.16 \cdot 10^{-1} \cdot IMmrKQg + 9.41 \cdot 10^{-3} \cdot IPMDKQg - 7.80 \cdot 10^{-2} \cdot IFMMKQg \quad \text{Eq.(2)}$$

The molecular descriptors used by the models, their calculated values, the estimated value of the octanol-water partition coefficient obtained with each model (\hat{Y}_{2D} - estimated octanol-water partition coefficient by the model with two descriptors, \hat{Y}_{4D} - estimated octanol-water partition coefficient by the model with four descriptors), and the values of residuals (defined as differences between measured octanol-water partition coefficient and estimated by the multi-varied model with two variables - $R_{\hat{Y}_{2D}}$ and by the multi-varied model with four variables $R_{\hat{Y}_{4D}}$, respectively) are presented in Table 2.

ET36/2005 – Et. Finală/2006 – Lucrare in extenso

Graphical representations of residuals obtained with the MDF-SPR models with two and four descriptors, respectively, are shown in Figure 1.

The statistical characteristics of the MDF-SPR models are presented in Table 3 and the quality characteristics of the regression models are shown in Table 4.

The plot of the estimated $\log K_{ow}$ by multi-varied MDF-SPR model with four descriptors versus measured $\log K_{ow}$ is presented in Figure 2.

A correlated correlation analysis was applied in order to verify the hypothesis that the correlation coefficient obtained by the model with four descriptors was not statistically different, at a significance level of 5%, compared with the correlation coefficient obtained by the model with two descriptors. The results are presented in Table 5.

The validation of the multi-varied MDF-SPR model with four descriptors was performed by spilling the sample of para-substituted phenols in training and test sets. The characteristics of the regression models and their performances are shown in Table 6. The following were included in Table 6: the coefficients of the regression models (using the generic model: $\hat{Y} = a_0 + a_1 \cdot isDDkGg + a_2 \cdot IMmrKQg + a_3 \cdot lPMDKQg + a_4 \cdot lFMMKQg$), the number of compound included in training (N_{otr}) and test (N_{ots}) sets, the multiple correlation coefficient of each training (r_{tr}) and test (r_{ts}) sample and associated 95% confidence intervals (95%CI r_{tr} - for training sets, and 95%CI r_{ts} - for test sets), Fisher parameter and its significance, at a significance level of 5%, for training (F_{tr}) and test (F_{ts}) models, and Fisher Z test of comparison between the correlation coefficient obtained in training set and the correlation coefficient obtained in corresponding test set ($Z_{trr-sts}$).

The estimation and prediction abilities of the multi-varied MDF-SPR model with four descriptors obtained in training versus test analysis, when the number of compounds in training set was equal with 2/3 from the total number of compounds, is presented in Figure 3.

DISCUSSIONS

The MDF-SPR methodology proved to be a useful method in estimation and prediction of the octanol-water partition coefficient for studied para-substituted phenols, this property being in relationship with complex information obtained from the compounds structure.

The best estimation and prediction abilities were obtained by the multi-varied MDF-SPR models with two and four descriptors (Eq.(1) and Eq.(2)).

The analysis of the MDF-SPR model with two descriptors (Eq.(1)) revealed that the octanol-water partition coefficient of studied para-substituted phenols was strongly related with molecular geometry (*isDDkGg*, *IMmrKQg*), being dependent on the partial charges

ET36/2005 – Et. Finală/2006 – Lucrare in extenso

(*IMmrKQg*) and group electronegativity (*isDDkGg*), directly related with the elastic force (*IMmrKQg*) and inverse related with the property potential (*isDDkGg*). For one descriptor (*isDDkGg*), its intercept had positive regression coefficients, while for other (*IMmrKQg*) had a negative one. The intercept of one descriptor (*isDDkGg*) had positive regression coefficients while the other (*IMmrKQg*) had a negative one.

The analysis of the performances of the model with two descriptors concluded that this was statistically significant in estimation as well as in prediction (see the squared correlation coefficient, adjusted values, and leave-one-out score, Table 3). Almost ninety percent of the octanol-water partition coefficient for studied para-substituted phenols can be explained by its linear relationship with the variation of *isDDkGg* and *IMmrKQg* descriptors (model with two descriptors, Table 3). The goodness-of-fit of the MDF-SPR model with two descriptors is sustained by the correlation coefficient, which is equal with 0.9457, its validity by the significance of the model and standard error, while its predictive abilities by the cross validation leave-one-out squared correlation coefficient, and by the Fisher parameter and its significance in leave-one-out analysis, which is less than 0.0001. The multi-varied MDF-SPR model with two descriptors proved to be a valid and stable model ($r^2_{cv-100} = 0.8660$; $r^2 - r^2_{cv-100} = 0.0284$).

The first thing which can be observed by analyzing the multi-varied MDF-SPR model with four descriptors (Eq.(2)) refers to the molecular descriptors used by the model: two of them are the descriptors used by the MDF-SPR model with two descriptors. The analysis of the molecular descriptors used by the multi-varied model with four descriptors suggests that the octanol-water partition coefficient of studied para-substituted phenols is strongly related with molecular geometry (*isDDkGg*, *IMmrKQg*, *IPMDKQg*, *IFMMKQg*), partial charges (*IMmrKQg*, *IPMDKQg*, *IFMMKQg*) and group electronegativity (*isDDkGg*), it is in relation with the elastic force (*IMmrKQg*, *IPMDKQg*, *IFMMKQg*), and inverse related with the property potential (*isDDkGg*).

The estimation abilities of the multi-varied MDF-SPR model with four descriptors are sustained by the value of the correlation coefficient ($r = 0.9890$, Table 3), confidence boundaries associated with the regression coefficients and probabilities associated with Student tests applied for the regression coefficients (for all coefficients less than 0.0001, see Table 4). Almost ninety-nine percent from the variation of the octanol-water partition coefficient of studied para-substituted phenols can be explained by its linear relationship with the variation of the four molecular descriptors used in the model (Eq.(2), Table 3). The value of the Fisher parameter ($F_{pred} = 189$) and its significance, which is less than 0.0001, support

ET36/2005 – Et. Finală/2006 – Lucrare in extenso

the prediction abilities of the model. The stability of the multi-varied MDF-SPR model with four descriptors is sustained by the values of difference between the correlation coefficient and the cross validation leave-one-out correlation score ($r^2 - r_{cv-100}^2 = 0.0100$), the value of the cross validation score being very close to the value of the squared correlation coefficient. The power of the model with four descriptors in prediction of octanol-water partition coefficient of studied para-substituted phenols is sustained by the absence of co-linearity between descriptors (see the squared correlation coefficients between pairs of descriptors, which is less than 0.33, with one exception, Table 3) and/or between $\log K_{ow}$ and descriptors (see the squared correlation coefficients in Table 4, which are less than 0.48).

The comparison between multi-varied MDF-SPR models with two and four descriptors, respectively, can be performed by analyzing the residuals and/or the correlation coefficients. As far as the residuals were concerned, their values obtained by the MDF-SPR model with two descriptors varied from -1.1771 to 1.1861 while the values obtained by the model with four descriptors varied from -0.8964 to 0.5717. The analysis of the absolute value of residuals obtained by the MDF-SPR models revealed that the minimum values were obtained in nineteen cases by the MDF-SPR model with four descriptors. The comparison between MDF-SPR models revealed that the model with four descriptors obtained a significantly greater correlation coefficient compared with the model with two descriptors ($p < 0.0001$, Table 5). The regression model with two, as well as the model with four descriptors, respects the specification of Hawkins D. M. [24] regarding the number of descriptors according with sample size.

The goodness-of-fit of the multi-varied MDF-SPR model with four descriptors and its internal predictivity was assessed in training versus test analysis. The analysis was performed by splitting the sample of compounds into training and test sets, the allocation of a compound into a set or into another being performed through randomization.

The analysis of the results concluded that, with two exceptions, the values of coefficients of the models in training sets did not exceeded the 95% confidence intervals of the multi-varied MDF-SPR model with four descriptors. With one exception, when the value was greater than the upper 95% confidence interval boundary, the correlation coefficients obtained in training and test sets did not exceed the 95% confidence intervals associated with the correlation coefficient of the multi-varied MDF-SPR model with four descriptors (see values in Table 3 and Table 6). As noted in Table 6, with one exception (for sample size in training set equal with 15), the correlation coefficients obtained in training sets were not statistical significant

ET36/2005 – Et. Finală/2006 – Lucrare in extenso

different, at a significance level of 5%, compared with the values obtained in test sets ($p > 0.05$, Table 6).

The multi-varied MDF-SPR model with four descriptors can be used in order to predict the octanol-water partition coefficient of para-substituted phenols without any experiments and measurements. By using the *MDF SPR Predictor* application [25], the property of a new para-substituted phenol can be obtained in a short time, provided that its structure is a *.hin file.

CONCLUSIONS

The octanol-water partition coefficient of para-substituted phenols proved to be strongly related with compounds geometry, partial charges and elastic force and in relation with group electronegativity and inverse related with the property potential.

The goodness-of-fit of the multi-varied MDF-SPR model with four descriptors and internal validation results sustain that the model is stable and valid. Future studies on new external para-substituted phenols are necessary in order to assess the robustness and predictivity of the multi-varied MDF-SPR model with four descriptors.

REFERENCES

- [1] Sangster, J. Octanol-Water Partition Coefficients: Fundamentals and Physical Chemistry; Wiley & Sons, Chichester, 1997.
- [2] Padmanabhan, J.; Parthasarathi, R.; Subramanian, V.; Chattaraj, P. K. *Bioorg Med Chem* 2006, 14(4), 1021-1028.
- [3] Kim, I.-H.; Morisseau, C.; Watanabe, T.; Hammock, B. D. *J Med Chem* 2004, 47(8), 2110-2122.
- [4] Dimitrov, S.; Koleva, Y.; Schultz, T. W.; Walker, J. D.; Mekenyan, O. *Environ Toxicol Chem*, 2004, 23(2), 463-470.
- [5] Puzyn, T.; Rostkowski, P.; Świeczkowski, A.; Jedrusiak, A.; Falandysz, J. *Chemosphere* 2006, 62(11), 1817-1828.
- [6] Leo, A. J. *Chem Rev* 1993, 93(4), 1281-1306.
- [7] Poulin, P.; Krishnan, K. *Toxicol Method* 1996, 6(3), 117-137.
- [8] Bodor, N.; Buchwald, P. *J Phys Chem B* 1997, 101(17), 3404-3412.
- [9] Pannier, A. K.; Brand, R. M.; Jones, D. D. *Pharm Res* 2003, 20(2), 143-148.
- [10] Zheng, G.; Huang, W. H.; Lu, X. H. *Anal Bioanal Chem* 2003, 376(5), 680-685.
- [11] Jäntschi, L. *Leonardo Electronic Journal of Practices and Technologies* 2005, 6, 76-98.
- [12] Jäntschi, L. *Leonardo Journal of Sciences* 2004, 4, 67-84.

ET36/2005 – Et. Finală/2006 – Lucrare in extenso

- [13] Jäntschi, L. Applied Medical Informatics 2004, 15, 48-55.
- [14] Jäntschi, L., Bolboacă, S. Leonardo Electronic Journal of Practices and Technologies 2006, 8, 71-86.
- [15] Jäntschi, L. Leonardo Journal of Sciences 2004, 5, 63-73.
- [16] Jäntschi, L.; Bolboacă, S. Leonardo Electronic Journal of Practices and Technologies 2005, 7, 55-102.
- [17] Schultz, T. W. Bull Environ Contam Toxicol 1987, 38(6), 994-999.
- [18] ***, HyperChem , Molecular Modelling System [online]; ©2003, Hypercube [cited 2005 Nov]. Available from: URL: <http://hyper.com/products/>
- [19] Jäntschi, L.; Katona, G.; Diudea, V. M. Commun Math Comput Chem (MATCH) 2000, 41, 151-188.
- [20] Diudea, M.; Gutman, I.; Jäntschi, L. Molecular Topology, 2nd Edition, Nova Science, Huntington: New York, 2002.
- [21] ***, Leave-one-out Analysis. ©2005, Virtual Library of Free Software [cited 2006 March]. Available from: URL: http://vl.academicdirect.org/molecular_topology/mdf_findings/loo/
- [22] Steiger, J. H. Psychol Bull 1980, 87, 245-251.
- [23] ***, Training vs. Test Experiment. ©2005, Virtual Library of Free Software [cited 2006 March]. Available from: URL: http://vl.academicdirect.org/molecular_topology/qsar_qspr_s/
- [24] Hawkins, D. M. J Chem Inf Comput Sci 2004, 44, 1-12.
- [25] ***, MDF SPR-SAR Predictor, © 2005, Virtual Library of Free Software [cited 2006 March]. Available from: URL: http://vl.academicdirect.org/molecular_topology/mdf_findings/sar

Molecular Descriptors Family on Structure-Activity Relationships: Modeling Herbicidal Activity of Substituted Triazines Class

Abstract

Herbicidal activity of a set of thirty 1,3,5-substituted-triazines were studied using an original structure-activity relationships approach. The cross-validation leave-one-out correlation score, the training vs. test analysis, and the model stability sustained the prediction ability of the best performing multi-varied model with four variables. The comparison with the previous reported model was performed by the use of correlated correlation analysis. The obtained multi-varied MDF-SAR model with four-descriptors shows that the herbicidal activity of 1,3,5-substituted-triazines is of geometrical and topological nature and is strongly depended on partial charges and number of directly bonded hydrogen's.

Keywords: Molecular Descriptor Family on Structure-Activity Relationships (MDF-SAR); Herbicidal Activity; Triazines; Multiple Linear Regressions (MLR);

Introduction

The Quantitative Structure-Activity Relationships is use today for finding the link between the activity and structure of chemical compounds in order to obtain new compounds with better properties, lowest expenses, and without time-consuming experiments [7].

The herbicidal activity of some 1,3,5-substituted-triazines, heterocyclic ring structures analogous to the six-members benzene ring with three carbon from positions 1, 3 and 5 replaced by nitrogen, were previous studied using orthogonalized molecular connectivity indices [8] and topological substituent descriptors [9]. The model and its statistical characteristics reported by Diudea & all [9] were:

[7] Peijnenburg WJGM. Structure-Activity Relationships for Biodegradation: A Critical Review, *Pure &Appl. Chem.*, 1994;66(9):1931-1941.

[8] Soskic M, Plavsic D, Trinajstic N. 2-Difluoromethylthio-4, 6-bis(monoalkylamino)-1, 3, 5-triazines as Inhibitors of Hill Reaction: A QSAR Study with Orthogonalized Descriptors. *J Chem Inf Comput Sci* 1996;36:146-150.

[9] Diudea VM, Jäntschi L, Pejov L. Topological Substituent Descriptors. *Leonardo Electronic Journal of Practices and Technologies* 2002;1:1-18.

ET36/2005 – Et. Finală/2006 – Lucrare in extenso

$$\text{Est } pI_{50} = 9.614 - 0.153 \cdot X_5 - 58.888 \cdot 1/V_5 - 2.430 \cdot 1/N_3$$

$$n = 30; r^2 = 0.9694; F = 274.3; r_{loo} = 0.9778$$

where X_5 = topological descriptor for substituent number 5, V_5 = fragmental volumes of the substituent in the position 5 (cm^3/mol); N_3 = total number of hydrogen's in the substituent 3; r^2 = squared correlation coefficient; F = Fisher parameter; and r_{loo} = squared correlation coefficient obtained by leave-one-out analysis.

According with the concepts of quantitatively correlating structure of compounds with their biological activities [10], starting from the successful results obtained by an original molecular descriptors family on structure-activity relationships (MDF-SAR) [11, 12, 13, 14, 15] herbicidal activity of a set of thirty 1,3,5-substituted-triazines was modeled by the use of MDF-SAR methodology and estimation and prediction abilities of the multi-varied models were analyzed.

Material and Method

The inhibition activity of thirty 1,3,5-substituted-triazines on *Chorella*, express as pI_{50} (the

[10] Kumar AD. Quantitative Structure-Activity Relationship (QSAR) Paradigm - Hansch Era to New Millennium. *Mini Rev Med Chem* 2001;1(2):187-195.

[11] Jäntschi L. Molecular Descriptors Family on Structure Activity Relationships 1. The review of Methodology. *Leonardo Electronic Journal of Practices and Technologies* 2005;6:76-98.

[12] Jäntschi L. Delphi Client - Server Implementation of Multiple Linear Regression Findings: a QSAR/QSPR Application. *Applied Medical Informatics* 2004;15:48-55.

[13] Bolboacă S, Jäntschi L. Molecular Descriptors Family on Structure Activity Relationships 3. Antituberculosic Activity of some Polyhydroxyxanthenes, *Leonardo Journal of Sciences* 2005;7:58-64.

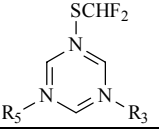
[14] Bolboacă S, Țigan Șt, Jäntschi L. Molecular Descriptors Family on Structure-Activity Relationships on anti-HIV-1 Potencies of HEPTA and TIBO Derivatives, Assa Reichert, George Mihalaș, Lăcrămioara Stoicu-Tivadar, Ștefan Schulz, Rolf Engelbrech (Eds.), *Proceedings of the European Federation for Medical Informatics Special Topic Conference*, April 6-8, 2006, Timișoara, Romania, p. 222-226.

[15] Jäntschi L, Ungureșan ML, Bolboacă SD. Integration of Complex Structural Information in Modeling of Inhibition Activity on Carbonic Anhydrase II of Substituted Disulfonamides. *Applied Medical Informatics* 2005;17:12-21.

ET36/2005 – Et. Finală/2006 – Lucrare in extenso

negative logarithm of concentration required for 50% inhibition of Hill reaction), was taken from a previous study [16]. The abbreviation of the compounds, the substituents in the positions 3 and 5 (R_3 and R_5 according with the below presented generic structure), the measured activity (pI_{50}), and previous estimated activity are in table 1.

Table 1. The substituent R_3 and R_5 of 1,3,5-triazines, measured (pI_{50}) and previous estimated (Est pI_{50} from [9]) activity

Abb.			pI_{50}	PrevEst pI_{50}
	R_3	R_5		
t_01	NH ₂	NH ₂	3.82	3.88
t_02	NHCH ₃	NH ₂	5.20	5.09
t_03	NHC ₂ H ₅	NH ₂	5.34	5.50
t_04	NH-i-C ₃ H ₇	NH ₂	5.83	5.70
t_05	NHCH ₃	NHCH ₃	6.01	6.10
t_06	NHC ₂ H ₅	NHCH ₃	6.39	6.51
t_07	NHC ₃ H ₇	NHCH ₃	6.75	6.71
t_08	NH-i-C ₃ H ₇	NHCH ₃	6.76	6.71
t_09	NHC ₄ H ₉	NHCH ₃	6.74	6.83
t_10	NH-s-C ₄ H ₉	NHCH ₃	6.76	6.83
t_11	NH-t-C ₄ H ₉	NHCH ₃	6.78	6.83
t_12	NHC ₅ H ₁₁	NHCH ₃	7.12	6.91
t_13	NHC ₃ H ₇	NHC ₂ H ₅	6.82	6.59
t_14	NHC ₃ H ₇	NHC ₂ H ₅	6.74	6.79
t_15	NH-i-C ₃ H ₇	NHC ₂ H ₅	6.89	6.79
t_16	NHC ₃ H ₇	NHC ₀ H ₅	6.95	6.91
t_17	NH-i-C ₄ H ₉	NHC ₂ H ₅	7.01	6.91
t_18	NH-s-C ₄ H ₉	NHC ₂ H ₅	6.87	6.91
t_19	NH-t-C ₄ H ₉	NHC ₂ H ₅	6.97	6.91
t_20	NHC ₅ H ₁₁	NHC ₂ H ₅	6.94	6.99
t_21	NHC ₆ H ₁₃	NHC ₂ H ₅	7.21	7.05
t_22	NHC ₇ H ₁₅	NHC ₂ H ₅	7.01	7.09
t_23	NHC ₈ H ₁₇	NHC ₂ H ₅	6.81	7.13
t_24	NHC ₃ H ₇	NHC ₃ H ₇	6.45	6.52
t_25	NHC ₃ H ₇	NH-i-C ₃ H ₇	6.75	6.65
t_26	NH-i-C ₃ H ₇	NH-i-C ₃ H ₇	6.75	6.65
t_27	NHC ₄ H ₉	NH-i-C ₃ H ₇	6.71	6.77
t_28	NH-s-C ₄ H ₉	NH-i-C ₃ H ₇	6.88	6.77
t_29	NH-t-C ₄ H ₉	NH-i-C ₃ H ₇	6.70	6.77
t_30	NHC ₅ H ₁₁	NH-i-C ₃ H ₇	6.69	6.85

[16] Morita K, Nagare T, Hayashi Y. Quantitative structure-activity relationships for herbicidal 2-Difluoromethylthio-4,6-bis(monoalkylamino)-1,3,5-triazines. Agric Biol Chem 1987;51:1955-1957.

ET36/2005 – Et. Finală/2006 – Lucrare in extenso

The steps applied in MDF-SAR modeling [17] were: (1) Sketch of the thirty 1,3,5-substituted-triazines compounds; (2) Creation of *triazines.txt* file (contain measured herbicidal activity for 1,3,5-substituted-triazines); (3) Generation of molecular descriptors family for studied 1,3,5-substituted-triazines; (4) Identification of MDF-SAR models; (5) Validation of MDF-SAR models; and (6) Analysis of the best performing MDF-SAR model in terms of estimation and prediction and comparison MDF-SAR model with previous reported QSAR.

The MDF SAR methodology was applied for modeling of herbicidal activity of the substituted triazines class in order to find a relationship between information obtained from the compounds structure and their herbicidal activity.

Many MDF members were obtained in step three. The names of each MDF member contain seven letters, which refers the characteristics used to build the descriptor. The 7th letter described the characteristics which consider the knowledge relative to the molecular geometry of the 1,3,5-substituted-triazines and its chemical model (topological vs. geometrical). The 6th letter of each descriptor is the atomic property; there were included six atomic properties (mass, charge, cardinality, electronegativity, group electronegativity, number of directly bonded hydrogen's). The others letters of descriptors described the interaction (the 5th letter) and the overlapping interaction (the 4th letter) of descriptors, the molecular fragmentation (the 3rd letter) and cumulated overall fragmental (the 2nd letter) descriptors and the linearization procedure applied (the 1st letter) to generate descriptors.

The selection of descriptors used in MDF-SAR modeling of 1,3,5-substituted-triazines took into consideration a total number of 74467 significantly different molecular descriptors. There was use a stepwise protocol for identification of best MDF-SAR models and all significantly different molecular descriptors were taken into consideration. First, a bias methodology was applied to the whole set of molecular descriptors at a significance level of 10^{-9} . The molecular descriptors were order by the obtained correlation coefficients between each descriptor and measured herbicidal activity and identical descriptors were eliminated. The next step was represented by including into the analysis pairs of two molecular descriptors. Because the best performing multi-varied MDF-SAR model with two variables did not obtained significantly better results comparing with the previous reported model, the modeling of the herbicidal activity of 1,3,5-substituted-triazines was move on by increasing with one the number of

[17] Jäntschi L. Molecular Descriptors Family on Structure Activity Relationships 1. The review of Methodology. Leonardo Electronic Journal of Practices and Technologies 2005;6:76-98.

ET36/2005 – Et. Finală/2006 – Lucrare in extenso

molecular descriptors until the obtained MDF-SAR multi-varied model had a significantly greater correlation coefficient.

A cross-validation leave-one-out procedure was applied in order to analyze the predictive performances of the MDF-SAR models. Each molecule from the whole set of thirty 1,3,5-substituted-triazines was deleted from the sample and the coefficients for multi-varied MDF-SAR models were rebuilt. The herbicidal activity of the deleted molecule was predicted by the use of the new MDF-SAR equation. At the end, the predicted activities for 1,3,5-substituted-triazines was correlated with measured activity and the standard error of estimated, Fisher parameter and its significance, as well as the cross-validation leave-one-out squared correlation coefficient were obtained.

The validation of the best performance multi-varied MDF-SAR model(s) was analyzed in training versus test sets by the use of *Training vs. Test* application [18]. The abilities of MDF-SAR models were analyzed by the use of a correlated correlation approach [19] in which correlation coefficients obtained by MDF-SAR models were compared with correlation coefficients obtained by previous reported QSAR [9].

Results

In MDF-SAR modeling of herbicidal activity of 1,3,5-substituted-triazines, from the total possible number of molecular descriptors (787968), 298462 proved to have real and distinct values. A number of 74467 MDF members were significantly different molecular descriptors. By the use of a series of MLR procedures, a pair of two descriptors, three descriptors and two pairs of two descriptors were correlated with measured herbicidal activity obtaining the best performing uni- and multi-varied models with two, three and respectively four descriptors. It was defined as best performing MDF-SAR model the one which obtained greater value for the squared correlation coefficient and for leave-one-out squared correlation coefficient. The best identified MDF-SAR models were:

- MDF-SAR model with one-variable:

$$\hat{Y}_{1v} = 7.47 - 4284.7 \cdot iSDRFHg$$

- MDF-SAR model with two-variables:

[18]***, Training vs. Test Experiment. ©2005, Virtual Library of Free Software [cited 2006 March]. Available from: URL: http://vl.academicdirect.org/molecular_topology/qsar_qspr_s/.

[19] Steiger JH. Tests for comparing elements of a correlation matrix. Psychological Bulletin 1980;87:245-251.

ET36/2005 – Et. Finală/2006 – Lucrare in extenso

$$\hat{Y}_{2v} = 5.52 - 8112.2 \cdot iSMMWHg + 194.35 \cdot iSMmEQt$$

- MDF-SAR model with three-variables:

$$\hat{Y}_{3v} = 1.74 - 9261 \cdot iSMMWHg + 10.34 \cdot iAMdEHg + 3.89 \cdot INDRJQg$$

- MDF-SAR model with four-variables:

$$\hat{Y}_{4v} = 5.75 + 199 \cdot iSMmEQt - 9010 \cdot iSMMWHg - 0.071 \cdot LADmkQt + 2.86 \cdot INPRJQg$$

where \hat{Y}_i ($i = 1, \dots, 4$) is the estimator of the herbicidal activity and $iSDRFHg$, $iSMmEQt$, $iSMMWHg$, $iAMdEHg$, $LADmkQt$, and $INPRJQg$ are MDF members.

The statistics associated with the MDF-SAR models are in table 2.

Table 2. Statistical characteristics of the best performing MDF SAR models

Characteristic (notation)	Value			
	4	3	2	1
Number of descriptors used in the model (n)	4	3	2	1
Correlation coefficient (r)	0.994	0.991	0.987	0.975
Squared correlation coefficient (r^2)	0.988	0.983	0.975	0.951
Adjusted squared correlation coefficient (r^2_{adj})	0.987	0.981	0.973	0.949
Standard error (S_{est})	0.081	0.096	0.114	0.156
Fisher parameter (F_{est})	537 ^{**}	501 ^{**}	533 ^{**}	549 ^{**}
Cross-validation leave-one-out correlation score ($r^2_{cv(loo)}$)	0.985	0.977	0.971	0.946
Fisher parameter for leave-one-out analysis (F_{pred})	409 ^{**}	361 ^{**}	449 ^{**}	488 ^{**}
Standard error - leave-one-out analysis (S_{loo})	0.092	0.113	0.124	0.165
Model stability ($r^2 - r^2_{cv(loo)}$)	0.003	0.006	0.004	0.006

^{**} p < 0.001

The squared correlation coefficient between each descriptor and measured herbicidal activity ($r^2(d, pI_{50})$), and the statistics of the MDF-SAR model with four-variables (express as coefficients of regression and associated 95% confidence interval (95%CI), standard error (StErr), t test parameter (t Stat) and its significance) are in table 3.

Table 3. The MDF-SAR model with four-variables: results of MLR

	$r^2(d, pI_{50})$	Coefficients [95%CI]	StErr	t Stat
Intercept	-	5.75 [5.28, 6.23]	0.23	24.88 [*]
iSMmEQt	0.6282	199 [161.27, 236.26]	18.21	10.92 [*]
iSMMWHg	0.9138	-9006 [-9712, -8300]	342.6	-26.28 [*]
LADmkQt	0.3224	-0.071 [-0.11, -0.03]	0.02	-4.05 [*]
INPRJQg	0.1599	2.86 [1.69, 4.03]	0.57	5.04 [*]

^{*} p < 0.05

The calculated values of the descriptors and the estimated herbicidal activity (\hat{Y}_{4v}) obtained for multi-varied MDF-SAR model with four descriptors, and the percent of residuals (R(%)) are in table 2. The residue represents the percent of the difference between the estimated activity with MDF-SAR model and measured activity reported to the measured activity.

Table 4. The calculated values of the descriptors, the herbicidal activity estimated by the MDF-SAR model with four descriptors ($a_1 \cdot iSMmEQt + a_2 \cdot iSMMWHg + a_3 \cdot LADmkQt + a_4 \cdot INPRJQg$) and the percent of residuals

Abb.	a_1	a_2	a_3	a_4	\bar{Y}_{4v}	R(%)
t_01	$2.04 \cdot 10^{-2}$	$7.04 \cdot 10^{-4}$	1.48	$1.63 \cdot 10^{-1}$	3.83	0.26
t_02	$1.98 \cdot 10^{-2}$	$5.21 \cdot 10^{-4}$	-1.51	$3.10 \cdot 10^{-2}$	5.20	0.00
t_03	$1.77 \cdot 10^{-2}$	$4.28 \cdot 10^{-4}$	1.83	$1.60 \cdot 10^{-2}$	5.33	-0.19
t_04	$1.57 \cdot 10^{-2}$	$3.24 \cdot 10^{-4}$	2.37	$1.88 \cdot 10^{-2}$	5.83	0.00
t_05	$1.95 \cdot 10^{-2}$	$4.02 \cdot 10^{-4}$	1.08	$1.30 \cdot 10^{-2}$	5.96	-0.83
t_06	$1.74 \cdot 10^{-2}$	$3.21 \cdot 10^{-4}$	0.73	$5.56 \cdot 10^{-2}$	6.43	0.63
t_07	$1.55 \cdot 10^{-2}$	$2.38 \cdot 10^{-4}$	2.59	$8.90 \cdot 10^{-2}$	6.76	0.15
t_08	$1.55 \cdot 10^{-2}$	$2.38 \cdot 10^{-4}$	1.75	$5.83 \cdot 10^{-2}$	6.72	-0.59
t_09	$1.55 \cdot 10^{-2}$	$2.08 \cdot 10^{-4}$	3.73	$1.78 \cdot 10^{-2}$	6.74	0.00
t_10	$1.38 \cdot 10^{-2}$	$1.70 \cdot 10^{-4}$	2.71	$9.73 \cdot 10^{-4}$	6.78	0.30
t_11	$1.37 \cdot 10^{-2}$	$1.64 \cdot 10^{-4}$	2.19	$3.81 \cdot 10^{-3}$	6.85	1.03
t_12	$1.47 \cdot 10^{-2}$	$1.60 \cdot 10^{-4}$	4.20	$4.58 \cdot 10^{-2}$	7.07	-0.70
t_13	$1.57 \cdot 10^{-2}$	$2.25 \cdot 10^{-4}$	2.07	$3.61 \cdot 10^{-2}$	6.80	-0.29
t_14	$1.48 \cdot 10^{-2}$	$1.86 \cdot 10^{-4}$	3.28	$3.08 \cdot 10^{-2}$	6.87	1.93
t_15	$1.33 \cdot 10^{-2}$	$1.54 \cdot 10^{-4}$	2.17	$2.66 \cdot 10^{-3}$	6.86	-0.44
t_16	$1.40 \cdot 10^{-2}$	$1.61 \cdot 10^{-4}$	3.99	$4.57 \cdot 10^{-3}$	6.83	-1.73
t_17	$1.37 \cdot 10^{-2}$	$1.55 \cdot 10^{-4}$	1.67	$1.73 \cdot 10^{-2}$	7.02	0.14
t_18	$1.33 \cdot 10^{-2}$	$1.44 \cdot 10^{-4}$	3.37	$9.13 \cdot 10^{-3}$	6.89	0.29
t_19	$1.25 \cdot 10^{-2}$	$1.28 \cdot 10^{-4}$	2.73	$2.13 \cdot 10^{-2}$	6.95	-0.29
t_20	$1.34 \cdot 10^{-2}$	$1.46 \cdot 10^{-4}$	4.37	$5.57 \cdot 10^{-2}$	6.95	0.14
t_21	$1.29 \cdot 10^{-2}$	$1.30 \cdot 10^{-4}$	4.67	$9.53 \cdot 10^{-2}$	7.08	-1.80
t_22	$1.24 \cdot 10^{-2}$	$1.18 \cdot 10^{-4}$	4.92	$7.80 \cdot 10^{-2}$	7.03	0.29
t_23	$1.20 \cdot 10^{-2}$	$1.07 \cdot 10^{-4}$	5.13	$5.03 \cdot 10^{-2}$	6.95	2.06
t_24	$1.40 \cdot 10^{-2}$	$1.85 \cdot 10^{-4}$	4.03	$2.78 \cdot 10^{-2}$	6.66	3.26
t_25	$1.33 \cdot 10^{-2}$	$1.71 \cdot 10^{-4}$	3.46	$2.09 \cdot 10^{-2}$	6.67	-1.19
t_26	$1.29 \cdot 10^{-2}$	$1.55 \cdot 10^{-4}$	2.53	$5.45 \cdot 10^{-3}$	6.76	0.15
t_27	$1.27 \cdot 10^{-2}$	$1.57 \cdot 10^{-4}$	4.09	$1.91 \cdot 10^{-2}$	6.62	-1.34
t_28	$1.23 \cdot 10^{-2}$	$1.40 \cdot 10^{-4}$	3.32	$2.40 \cdot 10^{-2}$	6.77	-1.60
t_29	$1.16 \cdot 10^{-2}$	$1.35 \cdot 10^{-4}$	2.76	$3.88 \cdot 10^{-2}$	6.76	0.90
t_30	$1.21 \cdot 10^{-2}$	$1.52 \cdot 10^{-4}$	4.43	$5.93 \cdot 10^{-2}$	6.65	-0.60

In order to evaluate the prediction ability of the MDF-SAR model with four-variables, the compounds were randomly split into two sets, training and test. A random routine pick out a specified number of compounds (n) from whole molecules (30), include them in the training set, and rebuild the MDF-SAR model. The prediction ability of the MDF-SAR model with four-descriptors was validated on 14 test sets (30 - n). The number of molecules in training sets varied from 10 to 23 (in test sets from 20 to 7) and the results are in table 5. In table 5, was used the following generic equation: $a_0 + a_1 \cdot iSMmEQt + a_2 \cdot iSMMWHg + a_3 \cdot LADmkQt + a_4 \cdot INPRJQg$, and the results are express as squared correlation coefficients (r^2_{tr} - for training set and r^2_{ts} - for test set), Fisher parameter and associated significance (less than 0.0001 if the value has one star (*)) and between 0.0001

ET36/2005 – Et. Finală/2006 – Lucrare in extenso

and 0.05 if the value has associated two stars (**)) for the MDF-SAR models, and the Fisher Z- test (FZ-test) which test the null hypothesis that there were not significant differences between correlation coefficient obtained in training set and the correlation coefficient obtained in the associated test set.

Table 5. The results of the training vs. test sets analysis using MDF-SAR model with four-descriptors

Coefficients					Training set			Test set			r _{tr} vs. r _{ts}
a ₀	a ₁	a ₂	a ₃	a ₄	No _{tr}	r ²	F _{tr}	No _{ts}	r ²	F _{ts}	F _{Z-test}
5.32	284.8	-12604	-0.12	4.009	10	0.897	11*	20	0.994	18**	3.22**
5.63	214.6	-9261	-0.084	2.562	11	0.989	138**	19	0.986	245**	0.28†
5.83	184.2	-8488	-0.054	1.678	12	0.995	353**	18	0.968	90**	2.22*
6.28	158.8	-8404	-0.096	2.154	13	0.991	230**	17	0.971	79**	1.43†
5.85	201.4	-9220	-0.084	1.763	14	0.987	177**	16	0.985	143**	0.18†
5.74	188.2	-8563	-0.043	3.16	15	0.989	230**	15	0.987	126**	0.21†
5.91	185.1	-8690	-0.075	2.096	16	0.99	280**	14	0.953	38**	1.91*
5.91	187.7	-8759	-0.074	2.033	17	0.977	125**	13	0.989	168**	0.90†
5.94	184.9	-8717	-0.086	2.359	18	0.995	612**	12	0.972	44**	2.06*
5.59	210.9	-9184	-0.069	3.137	19	0.993	516**	11	0.97	43**	1.69*
5.92	175.7	-8284	-0.066	2.813	20	0.908	37**	10	0.995	108**	3.29**
5.63	199.6	-8739	-0.041	2.538	21	0.988	330**	9	0.99	55**	0.19†
6.07	172.8	-8713	-0.075	3.329	22	0.981	220**	8	0.992	69*	0.87†
5.84	196.9	-9169	-0.09	3.462	23	0.989	411**	7	0.987	21*	0.15†

** p value < 0.001; * 0.05 < p-value < 0.001; † p > 0.05

Assessment of the MDF-SAR models were performed by the use of a correlated correlation analysis (Steiger's Z test), which took into consideration MDF SAR models with one-, two-, three- and respectively four-variables and compared them with previous reported QSAR [9]. The results of comparison are in table 6.

Table 6. The results of comparison between MDF-SAR models and previous reported QSAR

Characteristic	Value			
	4	3	2	1
r(pI ₅₀ , $\hat{Y}_{MDF-SAR}$)	0.994	0.991	0.987	0.975
r(pI ₅₀ , \hat{Y}_{QSAR})	0.985	0.985	0.985	0.985
r($\hat{Y}_{MDF-SAR}$, \hat{Y}_{QSAR})	0.988	0.981	0.986	0.979
Steiger's Z parameter	2.828*	1.563†	0.651†	-1.383†

* p < 0.05; † p > 0.05

Discussions

The herbicidal activity of a set of thirty 1,3,5-substituted-triazines was model using the MDF-SAR methodology.

Based on the values of squared correlation coefficients, the values of leave-one-out squared

ET36/2005 – Et. Finală/2006 – Lucrare in extenso

correlation coefficients and according with the correlated correlation analysis (see Table 2), it can be consider that the best performing MDF-SAR model is the model with four-variable (see Table 6).

The best performing MDF-SAR model with four-variable combines the geometrical shape (*iSMMWHg*, *INPRJQg*) as well as the topological shape of the molecules (*iSMmEQt*, *LADmkQt*), and as atomic property the partial charges of the molecule (*iSMmEQt*, *LADmkQt*, *INPRJQg*) and the number of directly bounded hydrogen's (*iSMMWHg*). According with the best performing multi-varied MDF-SAR model with four-variables, the herbicidal activity of studied compounds it is like to be of geometrical and topological nature and depend by the partial charges as well as by the directly bonded hydrogen's of the molecules. All compounds are alkyl-analogues and in these conditions if there are analyze exclusive the substituent, maybe the meaning of the partial charges will be questionable. But, in our study were took into consideration the whole molecule and its geometry as system, thus the meaning of partial charges can not influence the MDF-SAR model.

Ninety-nine percent of variation in herbicidal activity it is explainable by its linear relation with *iSMmEQt*, *iSMMWHg*, *LADmkQt*, and *INPRJQg* descriptors. Almost ninety-three percent of variation in herbicidal activity can be explainable by its linear relation with *iSMMWHg* descriptor and eighty-two percent by its linear relation with *iSMmEQt* descriptor. The values of squared correlation coefficient ($r^2 = 0.9885$) demonstrate the goodness of fit of the multi-varied MDFSAR model with four descriptors.

All coefficients associated with each molecular descriptor are significantly different by zero ($p < 0.001$) and had their role into MDF-SAR models. More, none of the confidence intervals associated with each of MDF-SAR model with four descriptors included the value equal with zero, which means that, each of them has a role in multi-varied model with four descriptors.

In the best performing MDF-SAR model with four descriptors, even if the 3rd and 4th descriptors seems to be insignificant, the MLR model which took into consideration just 1st and 2nd descriptors did not obtained a squared correlation coefficient significant different comparing with previous reported model (Steiger's Z test parameter = 0.651, $p > 0.05$).

The power of the MDF-SAR model with four descriptors in prediction of the herbicidal activity of 1,3,5-substituted-triazines is demonstrate by the cross-validation leave-one-out correlation score ($r^2_{cv(100)} = 0.9849$), procedure which did not take into consideration one molecule from the whole set. The stability of the best performing MDF-SAR model is give by the difference between the squared correlation coefficient and the cross-validation leave-one-out correlation score ($r^2 - r^2_{cv(100)} = 0.0035$).

ET36/2005 – Et. Finală/2006 – Lucrare in extenso

The prediction ability of the best performing MDF-SAR model with four descriptors is sustained by the results obtained in training vs. test analysis. The difference between leave-one-out procedure and training vs. test procedure is represented by the omission of more than one compound in training versus test analysis. It was not used an independent set for validation of the MDF-SAR model with four variables because the whole sample of thirty 1,3,5-substituted-triazines was used just for generating the list of descriptors. The algorithm of descriptors list generation is strictly based on the structure of the compounds. Any time the algorithm is used, for the same compound, the list of descriptors is the same; thus, splitting the compounds into training and test sets is not useful in descriptors list generation. The average of the squared correlation coefficients obtained for test sets (0.9814) it is not statistically grater comparing with the average of the squared correlation coefficients obtained for training sets (0.9765) and sustained the prediction ability of the model.

Comparing with the previous reported model [9] which use topological descriptors, the correlation coefficient obtained with the multi-varied MDF-SAR model with four descriptors is significantly greater (see table 6), sustaining its ability in prediction of herbicidal activity of 1,3,5-substituted-triazine compounds.

A software which allows to used the accumulated knowledge through learning of behavior on MDF-SAR models was developed in order to be apply to the new 1,3,5-subtituted-triazine compounds. The software is free to be use at:

http://vl.academicdirect.org/molecular_topology/mdf_findings/sar/

and is able to predict the activity of interest of a compound based on a choused class of compounds, choused model and on a *.hin file of compound of interest.

Thus, by using of the software from above address, the herbicidal activity of new 1,3,5-substituted-triazine can be calculated without any experiments. Unfortunately, in the stage in that the research was performed based on the obtained models, was proved that the model is useful just to be use to obtain new compounds. In the future research we intended to explore the physicochemical nature of each descriptor, but, as it results from the manual of the program these are very complex.

The future MDF-SAR study of herbicidal activity of a substituted triazines class will straighten on physicochemical properties of each descriptor and on mechanism of drug-descriptor interaction, in order to found the usefulness of MDF in exploring drug-action.

In conclusion, even if it is a time-consuming method, the MDF-SAR methodology gives a solution in predicting the herbicidal activity of 1,3,5-substituted-triazines providing a stable and performing multi-varied MDF-SAR model with four descriptors.

Online System for Molecular Descriptors Family on Structure-Activity Relationships: Assessment and Characterization of Biological Active Compounds

Synopsis

The aim of the paper is to present an open system which integrates the results obtained by utilization of an original approach called Molecular Descriptors Family on Structure-Activity/Property Relationships (MDF-SAR/SPR) applied on classes of chemical compounds and its usefulness as precursors of models elaboration for new biological active compounds.

The MDF-SAR/SPR methodology integrates the complex information obtained from compound's structure into unitary efficient models able to explain compounds activities/properties.

The methodology was applied on a number of thirty-one sets of molecules, twenty-one of them containing biological active compounds. The best subsets of molecular descriptors family members able to estimate and predict activity/property of interest were identified and were integrated into the system.

The abilities of the MDF-SAR/SPR models were compared with previous reported models by the use of correlated correlation analysis, which indicated that the MDF-SAR/SPR methodology is reliable.

The MDF-SAR/SPR methodology opens a new pathway in understanding the relationships between compound's structure and activity/property, in activity/property prediction, and in discovery, investigation and characterization of new chemical compounds, more competitive as costs and activity/property.

Keywords

Open System, Molecular Descriptors Family on Structure-Activity/Property Relationships (MDF-SAR/SPR), Biological Active Compounds

Introduction

Beginning with nineteenth century, characterization of activities and/or properties of chemical compound was done by applying of structure-activity relationships (QSAR) or quantitative structure-property (QSPR) methodologies, mathematical approaches of linking chemical structure and property/activity of chemical compounds in a quantitative manner [1].

ET36/2005 – Et. Finală/2006 – Lucrare in extenso

Observations relative to the relationship between activity/property and compounds structure were actually attained before the apparition of the QSAR/QSPR concepts. In 1868, Crum-Brown and Fraser stipulate the idea that the activity of a chemical is a function of its chemical composition and structure [2]. In 1893, Richet and Seancs showed for a set of organic molecules that the cytotoxicity is inverse related with water solubility [3]. Mayer suggests in 1899 that the narcotic action of a group of organic compounds is related with solubility into olive oil [4]. Ferguson introduced in 1939 a thermodynamic generalization to the correlation of depressant action with the relative saturation of volatile compounds in the vehicle in which they were administered [5]. Hammett [6] and Taft [7] put together the mechanistic basis of QSAR/QSPR development.

Ten years after defining of the QSAR/QSPR methods, these paradigms found their applicability in practice of agro-chemistry, pharmaceutical chemistry, toxicology and other chemistry related fields [8].

In QSAR/QSPR analysis, the electronic [9,10], hydrophobic [11,12], steric [13,14] and topologic [15,16] descriptors are most frequently used. Pure topological indices used in QSAR/QSPR analysis are Wiener index [17], Szeged index [18], and Cluj index [19,20].

The QSAR approach is used nowadays in drug investigations being seen as a useful tool in design of new compounds [21,22], in characterization of activity by the use of gene expression programming [23], and in analysis of the relationships between compounds structure and their biological activities [24,25].

An original approach called Molecular Descriptors Family on Structure- Activity/Property Relationships (MDF-SAR/SPR) was developed [26]. The MDF-SAR/SPR methodology, a unitary approach based on minimal complex knowledge obtained from the compound's structure, was applied on different classes of chemical compounds. Starting with the MDF-SAR/SPR models, an opens system was developed and assessed in order to provide a virtual experimental environment with applicability in analysis and characterization of compounds activities.

Material and Method

A number of thirty-one classes of chemical compounds were investigated with MDF-SAR/SPR methodology. Twenty-one out of thirty-one sets contained biological active compounds. Compounds abbreviation, the type of observed and/or measured activity/property, and compounds class are in table 1.

ET36/2005 – Et. Finală/2006 – Lucrare in extenso

Table 1. Set abbreviation, observed/measured activity/property, and class of compounds

Set name	Observed/Measured Activity/ Property	Compounds
IChr	retention chromatography index	organophosphorus herbicides
PCB_rrf	relative response factor	polychlorinated biphenyls
23159	octanol/water partition coefficients	polychlorinated biphenyls
23159e	octanol/water partition coefficients	polychlorinated biphenyls
RRC433_lkow	octanol/water partition coefficient	substituted phenols
PCB_lkow	octanol/water partition coefficient	polychlorinated biphenyls
36638	water activated carbon adsorption	organic compounds
MR10	molar refraction	cyclic organophosphorus
33504	boiling point	alkanes
PCB_rrt	relative retention time	polychlorinated biphenyls
Ta395	Cytotoxicity	quinolines
52730	Toxicity	alkyl metal compounds
RRC_lbr	Toxicity	para substituted phenols
23110	Toxicity	benzene derivates
23158	Toxicity	mono-substituted nitrobenzenes
23167	Toxicity	polychlorinated organic compounds
RRC_pka	relative toxicity	para substituted phenols
31572	irritation activity on eye	volatile organic compounds
Tox395	Mutagenicity	quinolines
41521	insecticidal activity	neonicotinoids
Triazines	herbicidal activity	substituted triazines
Dipeptides	inhibitory activity	dipeptides
52344	antioxidant efficacy	3-indolyl derivates
26449	antituberculosic activity	polyhydroxyxanthenes
23151	antimalarial activity	2,4-diamino-6-quinazoline sulfonamides
22583	anti-HIV-1 potencies	HEPTA and TIBO derivatives
19654	antiallergic activity	substituted N 4-methoxyphenyl benzamides
40846_1	inhibition activity on carbonic anhydrase I	substituted 1,3,4-thiadiazole- and 1,3,4-thiadiazoline-disulfonamides
40846_2	inhibition activity on carbonic anhydrase II	
40846_4	inhibitory activity on carbonic anhydrase IV	
3300	growth inhibitory activity	taxoids

The steps of molecular descriptors family on structure-activity/property relationships integrates: (1) the approaches of compounds preparing for molecular modeling, (2) the methodology of molecular descriptor family generation, (3) the methodology of finding the best performing MDF-SAR/SPR models, (4) the validation of the obtained MDF-SAR/SPR models, and (5) the comparison of the MDF-SAR/SPR models with previous reported models. Details regarding the MDF-SAR/SPR approach were previously published [26].

Starting from the previously experience in development of online systems [27,28], PHP (Hypertext Preprocessor), MySQL and Apache triad was used in creation of the open system. The reasons of chousing the triad are: (1) PHP is a server-side HTML embedded scripting language that supports dynamic web pages, freely available and used primarily but not just on Linux Web servers; (2) MySQL is a reliable and flexible open source relational database management system, and (3) Apache is an open source web server.

ET36/2005 – Et. Finală/2006 – Lucrare in extenso

The characteristics of the previous reported models and on the MDF-SAR/SPR models were summarized with Statistica software. The correlation coefficient obtained by the previous reported models was compared with the correlation coefficient obtained with MDF-SAR/SPR models by applying of the Fisher's Z test [29], at a significance level of 5%.

Results

The open system integrates six distinct programs useful in analysis and characterization of compounds activities/properties. The system is hosted by AcademicDirect domain being available at the following URL:

http://vl.academicdirect.org/molecular_topology/mdf_findings/

The named and the functions of the programs are in Table 2.

Table 2. MDF-SAR/SPR open system: programs and characteristics

Program	Functions
Browse/Query MDF-SARs/SPRs by set	
Browse	Display for a set of data the MDF-SAR/SPR equations, accompanied by the squared correlation coefficient, the number of descriptors, and the number of compounds.
Query	Display the followings characteristics of MDF-SAR/SPR investigations: the size of MDF, the MDF-SAR/SPR equations, the number of descriptors, the number of compounds, the values of descriptors in each model, the squared correlation coefficient, the leave-one-out score, the squared correlation coefficient between each descriptor and measured activity.
MDF (Demo) Calculator	
	Demo calculation of the Molecular Descriptors Family for a specified compound (a *.hin file) based on characteristics choused by the user.
MDF SAR Predictor	
	A set of previously obtained MDF-SAR/SPR models is considering the learning set. The activity of a new compound from the same class as learning compounds can be predicted based on its structure. A *.hin file with the structure of the compound of interest is necessary.
Leave-one-out Analysis	
	Based on the data resulted form MDF-SAR/SPR investigation, saved as *.html file, the program is able to compute the leave-one-out score and to display statistical characteristics of the estimated and predicted activity (number of descriptors used into the model, degree of freedom, standard error, standard deviation, squared correlation coefficient, Fisher parameter and associated significance). The program is able to work just with tabulated data (with labels on columns and rows) organized as followed: independent variables (first sets of columns, estimated dependent variable, measured/observed dependent variable, and predicted variable).
MDF Investigator	
	Display the characteristics of the sets of molecules which are in work. The administrator of the system is able to delete the MDF-SAR/SPR models which he considered not being at the level of imposed conditions.
Training vs. Test Experiment	
	Based on previously analyzed set of compounds, the experiment will randomly split the compounds into training and test sets (the user can choused the number of compounds in training set). The MDF-SAR/SPR equation is calculated on training set and applied on test set. The program display the molecular descriptors and associated values for compounds in training and test sets, the MDF-SAR/SPR equation, the squared correlation coefficient, the Fisher parameter and associated significance for both sets.

ET36/2005 – Et. Finală/2006 – Lucrare in extenso

The system assessment can be performed through analysis of the MDF-SAR/SPR estimation and prediction abilities. The summaries of characteristics of the previously reported results and of the MDF-SAR/SPR models are in table 3.

Characteristics of previously reported models obtained on biological active compounds and of MDF-SAR models, express as number of compounds (previously reported = n_{prev} , MDF-SAR = n_{MDF}), number of variables (previously reported = v_{prev} , MDF-SAR = v_{MDF}), squared correlation coefficient (previously reported = r^2_{prev} , MDF-SAR = r^2_{MDF}) and Fisher's Z parameter ($Z_{prev-MDF}$) of comparison between correlation coefficient of previous reported model and MDF-SAR model are in table 4.

Table 3. Statistical characteristics of the previously reported models and MDF-SAR/SPR models

Characteristic	Previously reported	MDF-SAR/SPR
Compounds with Activities and Properties		
Sample size		
Min	8	8
Max	73	209
Average [95%CI]	32.06 [26.20-37.92]	45.71 [24.87-66.55]
Number of variable		
Min	1	1
Max	7	5
Median	4	2
Mode	5	2
Squared correlation coefficient [95%CI]	0.86 [0.82-0.90]	0.90 [0.87-0.92]
Leave-one-out score [95%CI]	n.a.	0.88 [0.85-0.90]
Biological active compounds		
Sample size		
Min	8	8
Max	69	69
Average [95%CI]	27.59 [22.92-32.26]	29.90 [22.03-37.78]
Number of variable		
Min	1	1
Max	7	5
Median	4	2
Mode	5	2
Squared correlation coefficient [95%CI]	0.84 [0.80-0.88]	0.89 [0.87-0.92]
Leave-one-out score [95%CI]	n.a.	0.87 [0.84-0.90]

n.a. = not available

Table 4. Characteristics of previous reported models and MDF-SAR models

Set abb.	Previously reported model				MDF-SAR model				$Z_{prev-MDF}$
	n_{prev}	v_{prev}	r^2_{prev}	Ref.	n_{MDF}	v_{MDF}	r^2_{MDF}	Ref.	
40846_1	20	7	0.9170	[30]	40	4	0.9180	n.a.	0.021
40846_2	20	6	0.9020		40	4	0.9040	[31]	0.039
40846_4	20	4	0.8220		40	4	0.9200	[32]	0.068
RRC_lbr	30	2	0.9550	[33]	30	4	0.9739	n.a.	1.027
52344	8	4	0.9700	[34]	8	2	0.9998	[35]	3.972*
19654	23	3	0.8865	[36]	23	4	0.9978	[37]	6.329*
3300	35	5	0.9790	[38]	35	4	0.9665	n.a.	0.939
31572	24	4	0.9530	[39]	24	4	0.9583	n.a.	0.197

ET36/2005 – Et. Finală/2006 – Lucrare in extenso

23151	13	4	0.9850	[40]	16	3	0.9970	[41]	1.917 [*]
23158	40	5	0.8000	[42]	40	2	0.9510	n.a.	3.206 [*]
41521	8	5	0.9850	[43]	8	2	0.9990	[44]	2.144 [*]
Ta395	13	2	0.8700	[45]	15	2	0.9770	[46]	2.086 [*]
Tox395	13	2	0.8000		14	2	0.9570		1.86 [*]
Triazines	30	3	0.9700	[47]	30	4	0.9890	[48]	1.864 [*]
23167	27	3	0.9300	[49]	31	3	0.9390	n.a.	0.253
Dipeptides	58	2	0.7820	[50]	58	5	0.9250	n.a.	3.011 [*]
22583	57	5	0.8830	[51]	57	5	0.9180	[52]	0.97
23110	69	5	0.9000	[53]	69	5	0.9360	n.a.	1.339

number of compounds used by n_{prev} = previously reported model and n_{MDF} = MDF-SAR model;
number of variables used by v_{prev} = previously reported model and v_{MDF} = MDF-SAR model;
squared correlation coefficient of r_{prev}^2 = previously reported model and r_{MDF}^2 = MDF-SAR model;
 $Z_{\text{prev-MDF}}$ = Fisher's Z parameter of comparison between correlation coefficients;
^{*} $p < 0.05$; n.a. = not available

Discussion

The paper presented an open system on studying the relationships between activity/property and structure of chemical compounds by the use of molecular descriptors family on structure-activity/ property relationships (MDF-SAR/SPR).

Explosive development of information and communication technologies as well as computational technologies open new way in development of experimental research, providing devices and environments able to substitute the traditional experiments. The developed open system provides an environment of modeling the activity/property of chemical compounds assisted by a computer, an alternative free of risks procedure. The analysis of the system can be done through its advantages and disadvantages.

The advantages offered by the system are:

- Calculate based on information obtained strictly from the compounds structure the molecular descriptors family for every class of compounds;
- Identify the best performing models based on generated molecular descriptors family;
- Display a summary report of statistical characteristics of the best performing models;
- Apply methods of measures of goodness-of-fit, robustness and predictivity;
- Allows the user to visualize a demo of how the program calculate molecular descriptors family;
- Reproducibility: on a set of compounds the models will be identical any time the MDF-SAR analysis is repeat;
- Predict the activity/property of new compounds based from the same class as a previously analyzed class of compounds.

Note that the costs of virtual experiment are less comparing with real experiments. Also the experiment risks are withdrawn.

ET36/2005 – Et. Finală/2006 – Lucrare in extenso

As disadvantage, it must be mentioned that the process of generation the molecular descriptors family is time consuming, the time depending by the number of compounds, the power of calculus, and the computer performances. In order to use the system facilities, the user must to have browses computer skills, a computer connected to Internet being necessary. This can be consider at least for the researchers from developing countries a disadvantage of the system.

The analysis of the system can also be done through analysis of the MDF-SAR/SPR results. Analyzing the results obtained by MDF-SAR/SPR approach more observation can be done. First observation refers the number of compounds used by models. It is well known that discarding some molecules from the set and/or increasing the number of variables can lead to better model. Comparing with previously reported models, the MDF-SAR/SPR models always use the whole sample of compounds, the maximum number of compounds being on average almost three times greater in MDF-SAR/SPR models comparing with previously reported models (see table 3 and 4). The second observation refers the number of variables. Without any exceptions, the numbers of variables in MDF-SAR/SPR models are equal or less with up to two variables comparing with previously reported models, the median and the mode being two times less (see table 3 and 4). The third observation refers the squared correlation coefficient (see table 3). Looking on the results on whole sets of data, the difference between the averages of the squared correlation coefficient obtained with MDF-SAR/SPR and previously reported is of 0.04. A little higher difference is obtained on biological active compounds (0.05). The fourth observation refers the squared correlation coefficient obtained in leave-one-out analysis, this parameter being considered as a parameter of model predictivity. The 95% confidence intervals of squared correlation coefficient and of leave-one-out scores are overlapping, showing good predictive abilities.

Analyzing the results obtained by MDF-SAR on biologic active compounds, in the half of cases the correlation coefficient obtained by MDF-SAR model are statistically significant greater comparing with previously reported models (see table 4). In the other half of the cases, even if the correlation coefficient is not significant greater, some MDF-SAR models obtained same performances by using fewer variables (look at set 40846_1, 40846_2, 3300). It is well known that the fitting power of the model become greater by increasing the number of variables, being generally accepted that a regression model with ν descriptors for a sample size equal with n could be acceptable for validation only if the following criterion is satisfied: $n \geq 4 \cdot 5 \cdot \nu$ [54]. If there is applied the Hawkins criterion considering $4 \cdot \nu$ five out of eighteen previously reported models could not be considered valid and for the criterion of $5 \cdot \nu$ none of

ET36/2005 – Et. Finală/2006 – Lucrare in extenso

them. By applying the Hawkins criterion, even if it is use the 4- ν or 5- ν , all MDF-SAR models are acceptable for validation. On ensemble, it can be observed that the goodness-of-fit of the MDF-SAR models is very close to the ideal value (always greater than 0.9), and comparing with previous reported models, in half of the cases, the MDF-SAR approach provide better models.

The MDF-SAR/SPR proved to be reliable and valid. The results indicate that important information regarding compounds activity/property can be obtained by analyzing of compounds structure. Comparing with the experimental approach, the proposed online system provides a stable and valid alternative in studying of relationships between compounds structure and theirs activity/property.

The open system provide effective models which can be used in studying the activity of new compounds in real time, without any experiments, and with low costs, being necessary just building up as *.hin files the three dimensional structure of the new compound. The future development of the system will allow the access to exhaustive sets of compounds, opening a new pathway in study of their activities.

References

- [1] Hansch, C. *Acct Chem Res* 1969, 2, 232-239.
- [2] Crum-Brown, A.; Fraser, T. R. *Trans R Soc Edinburgh* 1868, 25, 151.
- [3] Richet, C.; Seancs, C. R. *Soc Biol Ses Fil* 1893, 9, 775.
- [4] Meyer, H. *Arch Exp Pathol Pharmacol* 1899, 42, 109.
- [5] Ferguson, J. *Proc R Soc London Ser B* 1939, 127, 387.
- [6] Hammett, L. P. *Chem Rev* 1935, 17, 125.
- [7] Taft, R. W. *J Am Chem Soc* 1952, 74, 3120.
- [8] Hansch, C.; Leo A. *Substituent Constants for Correlation Analysis in Chemistry and Biology*; John Wiley & Sons: New York, 1979.
- [9] Aptula, A.; Roberts, D. W.; Cronin, M. T. D.; Schultz, T. W. *Chem Res Toxicol* 2005, 18, 844-854.
- [10] Johnson-Restrepo, B.; Pacheco-Londoño, L.; Olivero-Verbel, J. *J Chem Inf Comput Sci* 2003, 43, 1513-1519.
- [11] Ponce, Y. M.; Marrero, R. M.; Castro, E. A.; De Armas, R. R.; Díaz, H. G.; Zaldivar, V. R.; Torrens, F. *Molecules* 2004, 9, 1124-1147.
- [12] Huibers, P. D. T.; Shah, D. O.; Katritzky, A. R. *J Colloid Interface Sci* 1997, 193, 132-136.

ET36/2005 – Et. Finală/2006 – Lucrare in extenso

- [13] Juvale, D. C.; Kadam, S. S.; Kulkarni, V. M. *Indian J Chem Sect A* 2006, 45, 194-201.
- [14] De Lima Ribeiro, F. A.; Ferreira, M. M. C. *J Mol Struct THEOCHEM* 2003, 663, 109-126.
- [15] Jin, A. Y.; Kohn, H.; Béguin, C.; Andurkar, S. V.; Stables, J. P.; Weaver, D. F. *Can J Chem* 2005, 83, 37-45.
- [16] Acevedo-Martínez, J.; Escalona-Arranz, J. C.; Villar-Rojas, A.; Téllez-Palmero, F.; Pérez-Rosés, R.; González, L., Carrasco-Velar, R. *J Chromatogr A* 2006, 1102, 38-244.
- [17] Wiener, H. J. *J Am Chem Soc* 1947, 69, 17-20.
- [18] Khadikar, P. V.; Deshpande, N. V.; Kale, P. P.; Dobrynin, A.; Gutman, I.; Domotor, G. J. *J Chem Inf Comput Sci* 1995, 35, 547-550.
- [19] Jäntschi, L.; Katona, G.; Diudea, M. V. *Commun Math Comput Chem (MATCH)* 2000, 41, 151-188.
- [20] Diudea, M.; Gutman, I.; Jäntschi, L. *Molecular Topology*; Nova Science, Huntington: New York, 2002: Chaper 6.
- [21] González, M. P.; Terán, C.; Teijeira, M.; Helguera, A. M. *Curr Med Chem* 2006, 13(19), 2253-2266.
- [22] Narasimhan, B.; Mourya, V.; Dhake, A. *Bioorg Med Chem Lett* 2006, 16(11), 3023-3029.
- [23] Si, H. Z.; Wang, T.; Zhang, K. J.; Hu, Z. D.; Fan, B. T. *Bioorg Med Chem* 2006, 14(14), 4834-4841.
- [24] Gallegos Saliner, A. *Curr Comput-Aided Drug Des* 2006, 2(2), 105-122.
- [25] Chang, C.; Swaan, P. W. *Eur. J Pharm Sci* 2006, 27(5), 411-424.
- [26] Jäntschi, L. *LEJPT* 2005, 6, 76-98.
- [27] Jäntschi, L. *LJS* 2002, 1, 31-52.
- [28] Bolboacă, S. D.; Jäntschi, L.; Deneş, C.; Achimaş Cadariu, A. *Roentgenologia & Radiologia*, 2005, XLIV(3), 189-193.
- [29] Steiger, J. H. *Psychol Bull* 1980, 87, 245-251.
- [30] Supuran, C. T.; Clare, B. W. *Eur J Med Chem* 1999, 34, 41-50.
- [31] Jäntschi, L.; Ungureşan, M. L.; Bolboacă, S. D. *Applied Medical Informatics* 2005, 17, 12-21.
- [32] Jäntschi, L.; Bolboacă, S. *Electron J Biomed* 2006, 2, In press.
- [33] Ivanciuc, O. *Revue Roumanian du Chimie* 1998, 43, 255-260.
- [34] Shertzer, G. H.; Tabor, M. W.; Hogan, I. T. D.; Brown, J. S.; Sainsbury, M. *Arch Toxicol* 1996, 70, 830-834.

ET36/2005 – Et. Finală/2006 – Lucrare in extenso

- [35] Bolboacă, S.; Filip, C.; Țigan, S.; Jäntschi, L. Clujul Medical 2006, LXXIX(2), 204-209.
- [36] Zhou, Y.-X.; Xu, L.; Wu, Y.-P., Liu, B.-L. Chemom Intell Lab Syst 1999, 45, 95-100.
- [37] Jäntschi, L.; Bolboacă, S. D. Clujul Medical 2006, LXXIX(3), In press.
- [38] Morita, H.; Gonda, A.; Wei, L.; Takeya, K.; Itokawa, H. Bioorg Med Chem Lett 1997, 7(18), 2387-2392.
- [39] Abraham, M. H.; Kumarsingh, R.; Cometto-Muniz, J. E.; Cain, W. S. Toxicol In Vitro 1998, 12, 201-207.
- [40] Agrawal, V. K.; Srivastava, R.; Khadikar, P. V. Bioorg Med Chem 2001, 9, 3287-3293.
- [41] Jäntschi, L.; Bolboacă, S. LJS 2006, 8, 77-88.
- [42] Agrawala, V. K.; Khadikar, P. V. Bioorg Med Chem 2001, 9, 3035-3040.
- [43] Hasegawa, K.; Arakawa, M.; Funatsu, K. Chemom Intell Lab Syst 1999, 47, 33-40.
- [44] Bolboacă, S.; Jäntschi, L. LJS 2005, 6, 78-85.
- [45] Smith, C. J.; Hansch, C.; Morton, M. J. Mutat Res 1997, 379, 167-175.
- [46] Jäntschi, L.; Bolboacă, S. The 10th Electronic Computational Chemistry Conference, April 2005, <http://eccc.monmouth.edu>
- [47] Diudea, M.; Jäntschi, L.; Pejov, L. LEJPT 2002, 1, 1-18.
- [48] Țigan, S.; Jäntschi, L.; Bolboacă, S.; Proceeding of the XXIII International Biometric Conference, Montreal, Canada, July 16-21, 2006.
- [49] Wei, D.; Zhang, A.; Wu, C.; Han, S.; Wang, L. Chemosphere 2001, 44, 1421-1428.
- [50] Opriș, D.; Diudea, M. V. SAR QSAR Environ Res 2001, 12, 159-179.
- [51] Castro, E. A.; Torrens, F.; Toropov, A. A.; Nesterov, I. V.; Nabiev, O. M. Mol Simul 2004, 30(10), 691-696.
- [52] Bolboacă, S., Țigan, Ș.; Jäntschi, L. Proceedings of the European Federation for Medical Informatics Special Topic Conference, April 6-8, 2006, Timișoara, Romania, p. 222-226.
- [53] Toporov, A. A.; Toporova, A. P. J Mol Struct Theochem 2002, 578, 129-134.
- [54] Hawkins, D. M. J Chem Inf Comput Sci 2004, 44, 1-12.

Molecular Descriptors Family on Structure Activity Relationships

5. Antimalarial Activity of 2,4-Diamino-6-Quinazoline Sulfonamide Derivates

Abstract

Antimalarial activity of sixteen 2,4-diamino-6-quinazoline sulfonamide derivates was modeled using an original methodology which assess the relationship between structure of compound and theirs activity. The results shows us that the antimalarial activity of studied 2,4-diamino-6-quinazoline sulfonamide compounds is alike topological and geometrical and is strongly dependent on partial change of the molecule. The ability in prediction with SAR models is sustained by the results obtained through cross-validation analysis and by the stability of the models. The SAR methodology gives us a real solution in structure-activity relationships investigation of 2,4-diamino-6-quinazoline sulfonamide compounds, obtaining better results by the use of two and/or three descriptors compared with the best performing previous reported model.

Keywords

Structure - Activity Relationships (SAR), Molecular Descriptors Family (MDF), Multiple Linear Regression (MLR), Antimalarial Activity, 2,4-diamino-6-quinazoline sulfonamide derivates

Background

The sulfonamides are sulfa-related group of antibiotics used in bacterial and some fungal infections, killing the bacteria and fungi by interfering with cell metabolism. Sulfonamides and its derivates, including the 2,4-diamino-6-quinazoline sulfonamides [20],

[20] Elslager E. F., Colbry N. L., Davoll J., Hutt M. P., Johnson J. L., Werbel L. M., *Folate antagonists. 22. Antimalarial and antibacterial effects of 2,4-diamino-6-quinazolinesulfonamides*, J. Med. Chem., 27(12), p. 1740-1743, 1984.

ET36/2005 – Et. Finală/2006 – Lucrare in extenso

have been used in medicine for their antimalarial properties [21]. To date, for 2,4-diamino-6-quinazoline sulfonamide derivatives, have been reported in specialty literature QSAR's models using electronic parameters, as energy of highest occupied molecular orbitals (EH), energy of lowest unoccupied molecular orbital (EL) and charge density (CD) [22] and topological properties (Wiener index - W , Szeged index - Sz , and indicator parameters, called dummy or de novo constants, which take two values – zero or one – and serve as indication of category or class membership - I_{p1} , I_{p2} and I_{p3}) [23]. Agrawal et al models the antimalarial activity of 2,4-diamino-6-quinazoline sulfonamide derivatives by the use of topological properties and obtained mono-, bi-, tri-, and tetra-parametric models. The models obtained previously are in table 1, indicating the regression equations, the square of correlation coefficient (r^2), and cross-validation values (r^2_{cv}) where were available.

Table 1. QSAR models for antimalarial activity of sulfonamide derivatives reported by Agrawal

No	QSAR model	r^2	r^2_{cv}
1	$6.9977-0.0032(\pm 9.9221 \cdot 10^{-4}) \cdot W$	0.4229	-
2	$6.8222-0.0019(\pm 6.3548 \cdot 10^{-4}) \cdot Sz$	0.3713	-
3	$9.6975-0.0045(\pm 0.0010) \cdot W-1.9814(\pm 0.08269) \cdot Ip_1$	0.5997	0.3325
4	$9.9246-0.0028(\pm 6.9413 \cdot 10^{-4}) \cdot Sz-2.1021(\pm 0.8938) \cdot Ip_1$	0.5590	-
5	$9.4033-0.0041(\pm 0.0010) \cdot W-2.1844(\pm 0.8013) \cdot Ip_1$ $- 1.0922(\pm 0.7280) \cdot Ip_2$	0.6629	0.5009
6	$9.4696-0.0026(\pm 7.2234 \cdot 10^{-4}) \cdot Sz-2.2182(\pm 0.8888) \cdot Ip_1- 0.9272(\pm 0.8072) \cdot Ip_2$	0.6027	0.3343
7	$9.0548-0.0019(\pm 7.5770 \cdot 10^{-4}) \cdot Sz-3.2559(\pm 0.9977) \cdot Ip_1$ $-2.6109(\pm 1.911) \cdot Ip_2-1.9345(\pm 1.0718) \cdot Ip_3$	0.6934	0.5579
8	$9.1679-0.0032(\pm 0.0010) \cdot W-3.1824(\pm 0.9143) \cdot Ip_1$ $-2.5978(\pm 1.0591) \cdot Ip_2- 1.7911(\pm 0.9807) \cdot Ip_3$	0.7414	0.6493

[21] Shao B. R., *A review of antimalarial drug pyronaridine*, Chin. Med. J. (Engl), 103, p. 428-434, 1990.

[22] Agrawal V. K., Sinha S., Bano S., Khadikar P. V., *QSAR studies on antimalarial 2,4-diamino-6-quinazoline sulfonamides*, Acta Microbiol Immunol Hung, 48(1), p. 17-26, 2001.

[23] Agrawal K. V., Srivastava R., Khadikar V. P., *QSAR Studies on Some Antimalarial Sulfonamides*, Bioorgan. Med. Chem., 9, p. 3287-3293, 2001.

ET36/2005 – Et. Finală/2006 – *Lucrare in extenso*

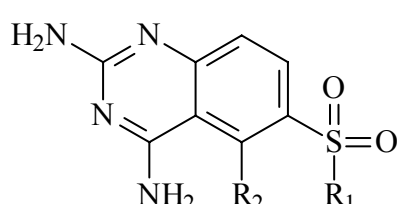
The aim of the research was to test the ability of SAR methodology in prediction of antimalarial activity of 2,4-diamino-6-quinazoline sulfonamide derivates and to compare the found models with previous reported QSARs.

Materials and Methods

Material and Pharmacology

Sixteen 2,4-diamino-6-quinazoline sulfonamide derivates was included into analysis. The planar structure of 2,4-diamino-6-quinazoline sulfonamide derivates, the substituents X and Y, and the measured antimalarial activity (Y_{aa}) are in table 2. Antimalarial activity used in the study was taken from the paper reported by Elslager et all [20], and is defined as difference between the average survival times (in days) of treated mice and the average survival times (in days) of control mice.

Table 2. Planar structure of 2,4-diamino-6-quinazoline sulfonamide derivates and measured antimalarial activity



No	R ₁	R ₂	Y _{aa}
mol_01	N(C ₂ H ₅) ₂	H	3.3
mol_02	N(CH ₂) ₅	Cl	2.3
mol_03	N(CH ₂ CH ₂ CH ₃) ₂	H	0.3
mol_04	N(CH ₂ CH ₂ OH) ₂	H	0.3
mol_05	N(CH ₃)CH (CH ₃) ₂	H	0.7
mol_06	N(CH ₃)CH ₂ CH ₂ N(C ₂ H ₅) ₂	H	0.1
mol_07	N(CH ₂) ₅	H	4.4
mol_08	N(CH ₂) ₄	H	5.0
mol_09	N[(CH ₂) ₂] ₂ O	H	4.7
mol_10	N[(CH ₂) ₂] ₂ S	H	2.5
mol_11	N[(CH ₂) ₂] ₂ NCH ₃	H	1.0
mol_12	N[(CH ₂) ₂] ₂ NC(=O)OC ₂ H ₅	H	0.2
mol_13	NH-C ₆ H ₄ -4Cl	H	0.7
mol_14	NH-C ₆ H ₄ -3Br	H	0.3
mol_15	NCH ₃ -C ₆ H ₄ -4Cl	H	0.3
mol_16	NCH ₃ -C ₆ H ₅	H	0.3

SAR modeling

The steps of molecular descriptors family on structure activity relationships modeling of antimalarial activity of 2,4-diamino-6-quinazoline sulfonamide derivates were [24]:

[24] Jäntschi L., *Molecular Descriptors Family on Structure Activity Relationships 1. The review of Methodology*, Leonardo Electronic Journal of Practices and Technologies, AcademicDirect, 6, p. 76-98, 2005.

ET36/2005 – Et. Finală/2006 – Lucrare in extenso

- Step 1: Sketch of 2,4-diamino-6-quinazoline sulfonamide compounds by the use of HyperChem software [25];
- Step 2: Create the file with measured antimalarial activity (Y_{aa}) of 2,4-diamino-6-quinazoline sulfonamide derivatives;
- Step 3: Generate the MDF members for the sixteen 2,4-diamino-6-quinazoline sulfonamide derivatives. Based on topological and geometrical representations of the sixteen 2,4-diamino-6-quinazoline sulfonamide, were calculated a total number of 289206 molecular descriptor. By applying significance selector to biases the values, and a significant difference value of 10^{-9} for mono-varied scores a number of 93362 molecular descriptors were found to be significant different and were included into analysis.
- Step 4: Find the SAR models for 2,4-diamino-6-quinazoline sulfonamide compounds. The criterion imposed in finding the SAR models was represented by the correlation coefficient and squared correlation coefficient most closed to the value equal with one.
- Step 5: Validation of the obtained SAR models were performed through computing the cross-validation leave-one-out correlation score [26], and the difference between this parameter and the squared correlation coefficient. The cross-validation leave-one-out correlation score was obtain after each compound from the whole set sixteen 2,4-diamino-6-quinazoline sulfonamide was deleted and the coefficients for the corresponding model (mono-, bi-, or tri-varied) were computed. The antimalaria activity of deleted compound was predicted by the use of new calculated equation (mono-, bi-, or tri-varied).
- Step 6: Analyze the selected SAR models and comparing them with previous reported model. The comparison between the SAR models and best performing previous reported QSAR was performed by applying the Steiger's Z-test.

Results

The best performing mono-, bi-, and tri-varied SAR models, together with associated statistics of regression analysis are in table 3.

[25] ***, HyperChem, Molecular Modelling System [Internet page], © 2003, Hypercube, Inc., available at: <http://hyper.com/products>.

[26] ***, Leave-one-out Analysis, © 2005, Virtual Library of Free Software, available at: http://vl.academicdirect.org/molecular_topology/mdf_findings/loo.

Table 3. SARs for antimalarial activity of 2,4-diamino-6-quinazoline sulfonamide derivatives with MDF members

No	SAR model	
	Characteristic	Notation and Value
1	<u>Mono-varied model:</u>	$\hat{Y}_{\text{mono-v}} = 3.26 \cdot 10^{-2} + 8.72 \cdot 10^{-5} \cdot \text{IsPmSQt}$
	Correlation coefficient	$r = 0.934$
	Squared correlation coefficient	$r^2 = 0.873$
	Adjusted squared correlation coefficient	$r^2_{\text{adj}} = 0.864$
	Standard error of estimated	$S_{\text{est}} = 0.659$
	Fisher parameter	$F_{\text{est}} = 96$
	Probability of wrong model	$p_{\text{est}}(\%) = 1.2 \cdot 10^{-5}$
	t parameter for intercept; p-values 95% probability CI _{int} [lower 95%; upper 95%]	$t_{\text{int}} = 0.140$; $p_{t_{\text{int}}} = 0.89$ $95\% \text{CI}_{\text{int}} = [-0.467; 0.533]$
	t parameter for IsPmSQt descriptor; p for t _{IsPmSQt} 95% probability CI _{IsPmSQt} [lower 95%; upper 95%]	$t_{\text{IsPmSQt}} = 9.802$; $p_{\text{IsPmSQt}} = 1.2 \cdot 10^{-7}$ $95\% \text{CI}_{\text{IsPmSQt}} = [6.81 \cdot 10^5; 10.63 \cdot 10^5]$
	Cross-validation leave-one-out (loo) score	$r^2_{\text{cv-loo}} = 0.840$
	Fisher parameter for loo analysis	$F_{\text{pred}} = 73$
	Probability of wrong model for loo analysis	$p_{\text{pred}}(\%) = 6.2 \cdot 10^{-7}$
	Standard error for leave-one-out analysis	$S_{\text{loo}} = 0.741$
	The difference between r^2 and $r^2_{\text{cv(loo)}}$	$r^2 - r^2_{\text{cv(loo)}} = 0.033$
2	<u>Bi-varied model:</u>	$\hat{Y}_{\text{bi-v}} = 4.81 \cdot 10^{-3} + 1.95 \cdot 10^{-5} \cdot \text{IsMMEQt} + 2.27 \cdot 10^{-7} \cdot \text{IIMMTQt}$
	Correlation coefficient	$r = 0.985$
	Squared correlation coefficient	$r^2 = 0.971$
	Adjusted squared correlation coefficient	$r^2_{\text{adj}} = 0.967$
	Standard error of estimated	$S_{\text{est}} = 0.324$
	Fisher parameter	$F_{\text{est}} = 220$
	Probability of wrong model	$p_{\text{est}}(\%) = 9.4 \cdot 10^{-9}$
	t parameter for intercept; p _{tint} 95% probability CI _{int} [lower 95%; upper 95%]	$t_{\text{int}} = 0.039$; $p_{t_{\text{int}}} = 0.969$ $95\% \text{CI}_{\text{int}} = [-0.261; 0.271]$
	t parameter for IsMMEQt descriptor; p _{IsMMEQt} 95% probability CI _{IsMMEQt} [lower 95%; upper 95%]	$t_{\text{IsMMEQt}} = 7.702$; $p_{\text{IsMMEQt}} = 3.4 \cdot 10^{-6}$ $95\% \text{CI}_{\text{IsMMEQt}} = [1.4 \cdot 10^5; 2.5 \cdot 10^5]$
	t parameter for IIMMTQt descriptor; p _{IIMMTQt} 95% probability CI _{IIMMTQt} [lower 95%; upper 95%]	$t_{\text{IIMMTQt}} = 17.74$; $p_{\text{IIMMTQt}} = 1.7 \cdot 10^{-10}$ $95\% \text{CI}_{\text{IIMMTQt}} = [2 \cdot 10^7; 2.5 \cdot 10^7]$
	Cross-validation leave-one-out (loo) score	$r^2_{\text{cv-loo}} = 0.961$
	Fisher parameter for loo analysis	$F_{\text{pred}} = 163$
	Probability of wrong model for loo analysis	$p_{\text{pred}}(\%) = 6.19 \cdot 10^{-8}$
	Standard error for leave-one-out analysis	$S_{\text{loo}} = 0.375$
	The difference between r^2 and $r^2_{\text{cv(loo)}}$	$r^2 - r^2_{\text{cv(loo)}} = 0.00958$
	The squared correlation coefficient between descriptor and measured antimalarial activity, and between descriptors	$r^2(\text{IsMMEQt}, Y_{\text{aa}}) = 0.277$ $r^2(\text{IIMMTQt}, Y_{\text{aa}}) = 0.840$ $r^2(\text{IsMMEQt}, \text{IIMMTQt}) = 0.035$
3	<u>Tri-varied model:</u>	$\hat{Y}_{\text{tri-v}} = -17.6 + 6.83 \cdot 10^{-8} \cdot \text{IsMMTQt} + 3.58 \cdot 10^{-1} \cdot \text{LsMrKQg} - 8.47 \cdot 10^{-1} \cdot \text{LsDMTQt}$
	Correlation coefficient	$r = 0.998$
	Squared correlation coefficient	$r^2 = 0.997$
	Adjusted squared correlation coefficient	$r^2_{\text{adj}} = 0.996$
	Standard error of estimated	$S_{\text{est}} = 0.1059$

ET36/2005 – Et. Finală/2006 – Lucrare in extenso

Fisher parameter	$F_{est} = 1415$
Probability of wrong model	$p_{est}(\%) = 1.4 \cdot 10^{-13}$
t parameter for intercept; p_{tint} 95% probability CI_{int} [lower 95%; upper 95%]	$t_{int} = -14.86$; $p_{tint} = 4.32 \cdot 10^{-9}$ $95\%CI_{int} = [-20.23; -15.05]$
t parameter for IsMMTQt descriptor; $p_{IsMMTQt}$ 95% probability $CI_{IsMMTQt}$ [lower 95%; upper 95%]	$t_{IsMMTQt} = 47.03$; $p_{IsMMTQt} = 5.58 \cdot 10^{-15}$ $95\%CI_{IsMMTQt} = [6.5 \cdot 10^8; 7.1 \cdot 10^8]$
t parameter for LsMrKQg descriptor; $p_{LsMrKQg}$ 95% probability $CI_{LsMrKQg}$ [lower 95%; upper 95%]	$t_{LsMrKQg} = 10.50$; $p_{LsMrKQg} = 2.09 \cdot 10^{-7}$ $95\%CI_{LsMrKQg} = [0.28; 0.43]$
t parameter for lsDMTQt descriptor; $p_{lsDMTQt}$ 95% probability $CI_{lsDMTQt}$ [lower 95%; upper 95%]	$t_{lsDMTQt} = -17.07$; $p_{lsDMTQt} = 8.8 \cdot 10^{-10}$ $95\%CI_{lsDMTQt} = [-0.95; -0.74]$
Cross-validation leave-one-out (loo) score	$r^2_{cv-loo} = 0.9959$
Fisher parameter for loo analysis	$F_{pred} = 970$
Probability of wrong model for loo analysis	$p_{pred}(\%) = 1.4 \cdot 10^{-12}$
Standard error for leave-one-out analysis	$S_{loo} = 0.1279$
The difference between r^2 and $r^2_{cv(loo)}$	$r^2 - r^2_{cv(loo)} = 0.0013$
The squared correlation coefficient between descriptor and measured antimalarial activity, and between descriptors	$r^2(IsMMTQt, Y_{maa}) = 0.8448$ $r^2(LsMrKQg, Y_{maa}) = 0.1556$ $r^2(lsDMTQt, Y_{maa}) = 0.2493$ $r^2(IsMMTQt, LsMrKQg) = 0.0135$ $r^2(IsMMTQt, lsDMTQt) = 0.6140$ $r^2(LsMrKQg, lsDMTQt) = 0.0242$

The list of descriptors and associated values used in mono-, bi-, and tri-varied models and estimated antimalarial activity (\hat{Y}) are in table 4.

Graphical representations of the antimalarial activity of 2,4-diamino-6-quinazoline sulfonamide derivates, obtained from structure for mono-, bi-, and tri-varied models vs. measured ones are in figures 1 to 3.

Assessment of the MDF SAR model was performed by applying a correlated correlation analysis, which took into consideration mono-, bi-, and tri-varied SAR models and compared them with the best performing (model with four variables, $r^2 = 0.7414$, $r^2_{cv} = 0.6493$) previous reported model [23] by the use of Steiger's Z test. The results of comparison are in table 5.

Table 4. Descriptors used in MDF SAR models, theirs values and estimated antimalarial activities

Mol	Mono-varied		Bi-varied			Tri-varied			
	IsPmSQt	\hat{Y}_{mono}	IsMMEQt	IIMMTQt	\hat{Y}_{bi}	IsMMTQt	LsMrKQg	lsDMTQt	\hat{Y}_{tri}
mol_01	$2.23 \cdot 10^{-6}$	1.98	$-2.01 \cdot 10^{-6}$	$1.53 \cdot 10^{-7}$	3.07	$7.63 \cdot 10^{-9}$	$-4.03 \cdot 10^0$	$-2.02 \cdot 10^1$	3.24
mol_02	$2.92 \cdot 10^{-6}$	2.58	$-1.46 \cdot 10^{-7}$	$6.98 \cdot 10^{-8}$	1.56	$3.17 \cdot 10^{-9}$	$-4.49 \cdot 10^0$	$-2.28 \cdot 10^1$	2.23
mol_03	$2.11 \cdot 10^{-7}$	0.22	$-1.12 \cdot 10^{-6}$	$1.57 \cdot 10^{-8}$	0.14	$7.13 \cdot 10^{-10}$	$-3.56 \cdot 10^0$	$-2.21 \cdot 10^1$	0.29
mol_04	$2.11 \cdot 10^{-7}$	0.22	$-1.12 \cdot 10^{-6}$	$1.57 \cdot 10^{-8}$	0.14	$7.13 \cdot 10^{-10}$	$-3.60 \cdot 10^0$	$-2.21 \cdot 10^1$	0.28
mol_05	$6.62 \cdot 10^{-7}$	0.61	$-1.28 \cdot 10^{-6}$	$6.60 \cdot 10^{-8}$	1.25	$3.30 \cdot 10^{-9}$	$-4.28 \cdot 10^0$	$-2.08 \cdot 10^1$	0.73
mol_06	$6.94 \cdot 10^{-7}$	0.64	$-6.59 \cdot 10^{-7}$	$1.94 \cdot 10^{-8}$	0.32	$8.08 \cdot 10^{-10}$	$-4.47 \cdot 10^0$	$-2.22 \cdot 10^1$	0.14
mol_07	$3.54 \cdot 10^{-6}$	3.12	$9.86 \cdot 10^{-6}$	$1.05 \cdot 10^{-7}$	4.31	$5.00 \cdot 10^{-9}$	$-1.45 \cdot 10^0$	$-2.26 \cdot 10^1$	4.40
mol_08	$5.91 \cdot 10^{-6}$	5.19	$4.82 \cdot 10^{-6}$	$1.73 \cdot 10^{-7}$	4.87	$8.65 \cdot 10^{-9}$	$-3.96 \cdot 10^0$	$-2.14 \cdot 10^1$	4.95

ET36/2005 – Et. Finală/2006 – Lucrare in extenso

mol_09	$5.18 \cdot 10^{-6}$	4.55	$-1.76 \cdot 10^{-6}$	$2.31 \cdot 10^{-7}$	4.89	$1.10 \cdot 10^{-8}$	$-4.69 \cdot 10^0$	$-1.96 \cdot 10^1$	4.76
mol_10	$4.20 \cdot 10^{-6}$	3.69	$1.39 \cdot 10^{-6}$	$1.15 \cdot 10^{-7}$	2.89	$5.49 \cdot 10^{-9}$	$-4.37 \cdot 10^0$	$-2.13 \cdot 10^1$	2.60
mol_11	$7.38 \cdot 10^{-7}$	0.68	$-5.86 \cdot 10^{-6}$	$8.47 \cdot 10^{-8}$	0.78	$3.85 \cdot 10^{-9}$	$-5.24 \cdot 10^0$	$-2.10 \cdot 10^1$	0.93
mol_12	$1.17 \cdot 10^{-7}$	0.13	$-4.82 \cdot 10^{-7}$	$1.28 \cdot 10^{-8}$	0.20	$4.92 \cdot 10^{-10}$	$-4.41 \cdot 10^0$	$-2.24 \cdot 10^1$	0.07
mol_13	$1.02 \cdot 10^{-6}$	0.93	$-5.55 \cdot 10^{-7}$	$4.25 \cdot 10^{-8}$	0.86	$1.85 \cdot 10^{-9}$	$-4.68 \cdot 10^0$	$-2.20 \cdot 10^1$	0.55
mol_14	$1.86 \cdot 10^{-7}$	0.20	$-5.58 \cdot 10^{-7}$	$1.22 \cdot 10^{-8}$	0.17	$5.31 \cdot 10^{-10}$	$-5.03 \cdot 10^0$	$-2.28 \cdot 10^1$	0.26
mol_15	$7.33 \cdot 10^{-7}$	0.67	$-1.24 \cdot 10^{-7}$	$2.33 \cdot 10^{-8}$	0.51	$9.70 \cdot 10^{-10}$	$-4.13 \cdot 10^0$	$-2.24 \cdot 10^1$	0.51
mol_16	$1.12 \cdot 10^{-6}$	1.01	$-3.72 \cdot 10^{-7}$	$2.27 \cdot 10^{-8}$	0.45	$9.88 \cdot 10^{-10}$	$-5.09 \cdot 10^0$	$-2.27 \cdot 10^1$	0.43

Table 5. The results of comparison obtained and best performing previous reported models

Characteristic	Values		
	3	2	1
Number of descriptors used in MDF SAR model	3	2	1
$r(Y_{aa}, \hat{Y}_{MDF\ SAR})$	0.9986	0.9856	0.9342
$r(Y_{aa}, \hat{Y}_{Previous})$	0.8598	0.8598	0.8598
$r(\hat{Y}_{MDF\ SAR}, \hat{Y}_{Previous})$	0.8641	0.8624	0.8139
Steiger's Z test parameter	7.8891	3.9686	1.3229
$p_{Steiger's\ Z}$ (%)	$1.5 \cdot 10^{-13}$	$3.6 \cdot 10^{-3}$	9.2926

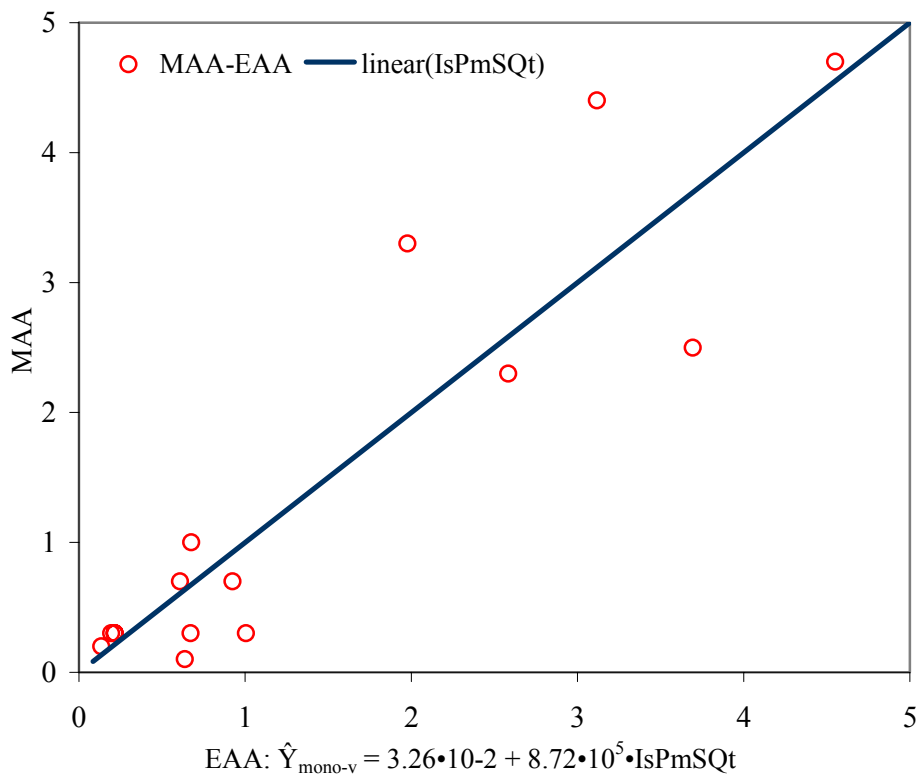


Figure 1. Measured antimalarial activity (MAA) vs. estimated (EAA) with mono-varied model

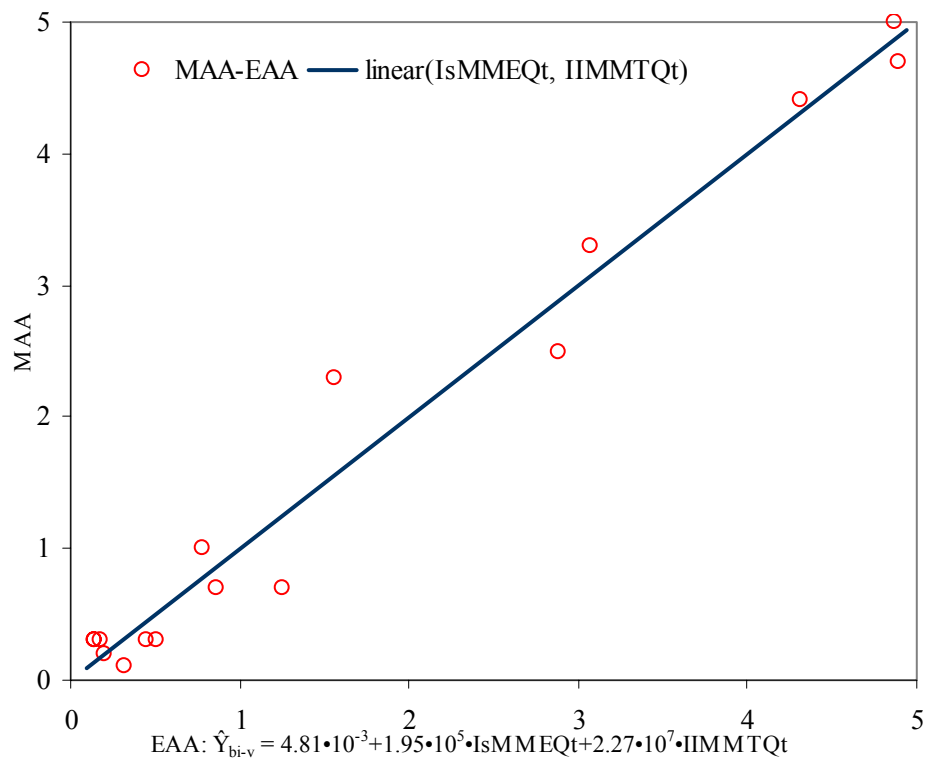


Figure 2. Measured antimalarial activity vs. estimated with bi-varied model

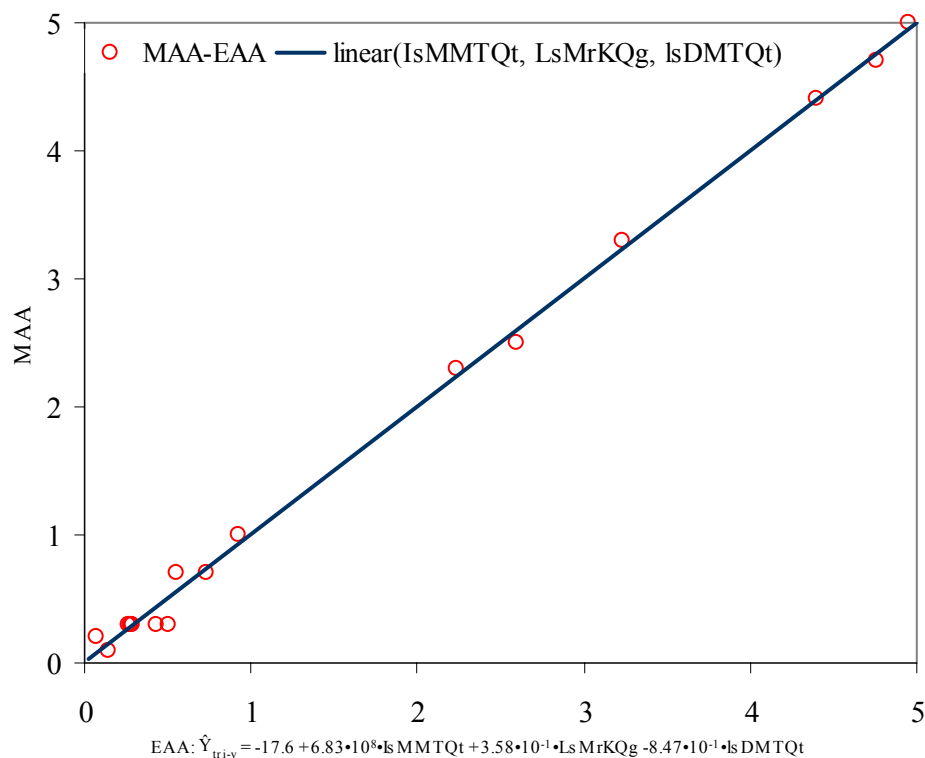


Figure 3. Measured antimalarial activity vs. estimated with tri-varied model

Discussions

Antimalarial activity of sixteen 2,4-diamino-6-quinazoline sulfonamide derivatives was modeled by the use of an original methodology which take into consideration the structure of the compound and try to explain the interest activity. Applying the MDF SAR methodology, three models, one mono-varied, one bi-varied and one-tri-varied prove to obtained performances in antimalarial activity prediction. All presented SAR models are statistically significant at a significance level less than 0.001. The mono-varied SAR model use a descriptor that take into consideration the topology of molecule (*IsPmSQt*) and the partial change as atomic property (*IsPmSQt*). Almost 87 percent of variation in antimalarial activity of 2,4-diamino-6-quinazoline sulfonamide derivatives can be explainable by its linear relation with *IsPmSQt* descriptor. The mono-varied model is significant different by the best performing four-varied model previous reported at a significance level equal with 0.09. As the mono-varied SAR model, the bi-varied one took into consideration the topology of molecule (*IsMMEQt*, *IIMMTQt*) as well as partial change as atomic property (*IsMMEQt*, *IIMMTQt*). All coefficients of the bi-varied equation are significantly differed by zero, except the intercept of the slop. The performance of the bi-varied SAR model is sustained by the correlation coefficient and the squared of the correlation coefficient. Ninety-seven percent of variation in antimalarial activity of 2,4-diamino-6-quinazoline sulfonamide derivatives can be explainable by its linear relation with *IsMMEQt*, *IIMMTQt* descriptors. The stability of the bi-varied model is proved by the very lower value of the differences between squared correlation coefficient and cross-validation leave-on-out squared correlation coefficient. The cross-validation leave-one-out score ($r^2_{cv-100} = 0.961$) sustain the stability of the bi-varied SAR model. Looking at the values of the squared correlation coefficient between descriptors and measured antimalarial activity it can be observed that there is no correlation between *IsMMEQt* descriptor and measured antimalarial activity but there is a strong correlation between *IIMMTQt* descriptor and antimalarial activity. Even if the correlation is strong, the *IIMMTQt* is not the one that obtained best performances in terms of squared correlation coefficient and cross-validation leave-one-out score in mono-varied SAR model. It could not be observed a significant correlation between descriptors of the bi-varied model ($r^2(\text{IsMMEQt}, \text{IIMMTQt}) = 0.035$). The bi-varied SAR model obtained a correlation coefficient significantly greater compared with the previous reported four-varied model at a significance level equal with $3.6 \cdot 10^{-3} \%$. Note that, it is possible to obtained useful information about antimalarial activity of 2,4-diamino-6-quinazoline sulfonamide derivatives

ET36/2005 – Et. Finală/2006 – Lucrare in extenso

with a bi-varied model instead of a model with four variable. Looking at the bi-varied model, we can say that the antimalarial activity is of molecular topology and depend on partial change of molecule. Looking at the cross-validation leave-one-out score, we can say that the tri-varied model is the best performing SAR model. Ninety-nine percent of variation in antimalarial activity of 2,4-diamino-6-quinazoline sulfonamide derivates can be explainable by its linear relation with IsMMTQt, LsMrKQg, and IsDMTQt descriptors. Two descriptors (IsMMTQt, and IsDMTQt) take into consideration the topology of the molecule while another one (LsMrKQg) the molecular geometry. All three descriptors (IsMMTQt, LsMrKQg, IsDMTQt) take into consideration the partial change of the molecule. The values of squared correlation coefficient ($r^2 = 0.997$) demonstrate the goodness of fit of the tri-varied MDF SAR model. The power of the tri-varied model in prediction of the antimalarial activity of 2,4-diamino-6-quinazoline sulfonamide compounds is demonstrate by the cross-validation leave-one-out correlation score ($r^2_{cv(100)} = 0.9959$), procedure which did not take into consideration one molecule from the whole set. The stability of the best performing tri-varied MDF SAR model is give by the difference between the squared correlation coefficient and the cross-validation leave-one-out correlation score ($r^2 - r^2_{cv(100)} = 0.0013$). Looking at the tri-varied SAR model we can say that the antimalarial activity of 2,4-diamino-6-quinazoline sulfonamide derivates is alike topological and geometrical and depend by the partial change of molecule. Looking at the correlated correlations analysis results, it can be observed that the tri-varied SAR model obtained a significantly greater correlation coefficient compared with the previous reported four-varied model, at a significance level equal with $1.5 \cdot 10^{-13}$ %. Starting with the knowledge learned from the studied set of 2,4-diamino-6-quinazoline sulfonamide compounds, antimalarial activity of new compound from the same class can be predict by the use of an original software, which is available at the following address:

http://vl.academicdirect.org/molecular_topology/mdf_findings/sar/

Thus, the software id able to predict the antimalarial activity of new 2,4-diamino-6-quinazoline sulfonamide compounds with low costs.

Conclusions

Antimalarial activity of the studied 2,4-diamino-6-quinazoline sulfonamide compounds is alike to be by topological and geometrical nature and is strongly dependent by partial change. The MDF SAR methodology is a real solution in predicting antimalarial

ET36/2005 – Et. Finală/2006 – Lucrare in extenso

activity of 2,4-diamino-6-quinazoline sulfonamide compounds and could be use in developing of new 2,4-diamino-6-quinazoline sulfonamide compounds with antimalarial properties.

Even if using of MDF in QSAR modeling is time consuming, it has doubtless advantages, such as better QSAR of antimalarial activity of 2,4-diamino-6-quinazoline sulfonamide derivates and a much closer structure activity explanation.

Pearson versus Spearman, Kendall's Tau Correlation Analysis on Structure-Activity Relationships of Biologic Active Compounds

Abstract

A sample of sixty-seven pyrimidine derivatives with inhibitory activity on *E. coli* dihydrofolate reductase (DHFR) was studied by the use of molecular descriptors family on structure-activity relationships. Starting from the results obtained by applying of MDF-SAR methodology on pyrimidine derivatives and from the assumption that the measured activity (compounds' inhibitory activity) of a biologically active compounds is a semi-quantitative outcome (can be related with the type of equipment used, the researchers, the chemical used, etc.), the abilities of Pearson, Spearman, Kendall's, and Gamma correlation coefficients in analysis of estimated toxicity were studied and are presented.

Keywords

Multiple linear regressions, Correlation coefficients, Molecular Descriptors Family on Structure-Activity Relationships (MDF-SAR)

Introduction

QSAR (Quantitative Structure-Activity Relationships) is an approach which is able to indicate for a given compound or a class of compounds which feature of structure characteristics is correlated with its activity [27]. In QSAR analysis were proposed several approaches for development. Simple and multiple linear regressions is one of the more successful techniques use by many researcher in construct of QSAR models [28-30].

[27] Rogers D., Hopfinger A. J., *Application of Genetic Function Approximation to Quantitative Structure-Activity Relationships and Quantitative Structure-Property Relationships*, J. Chem. Inf. Comput. Sci. 34, 1994, p. 854-866.

[28] Hansch C., Leo A., Stephen R., Eds. Heller, *Exploring QSAR, Fundamentals and Applications in Chemistry and Biology*, ACS professional Reference Book., American Chemical Society, Washington, D.C., 1995.

ET36/2005 – Et. Finală/2006 – Lucrare in extenso

Correlation coefficient is a simple statistical measure of relationship between one dependent and one or more than one independent variables and it is use as a measure of the statistical fit of a regression based model in QSAR [31]. Its squared value (the coefficient of determination) it is most frequently used parameter as a measure of the goodness-of-fit of the model [32-36].

A new approach of molecular descriptors family on structure-activity relationships (MDF-SAR) was developed [37], and proved its usefulness in estimation and prediction of:

[29] Zahouily M., Lazar M., Elmakssoudi A., Rakik J., Elaychi S., Rayadh A., *QSAR for anti-malarial activity of 2-aziridinyl and 2,3-bis(aziridinyl)-1,4-naphthoquinonyl sulfonate and acylate derivatives*, J Mol Model 12(4), 2006, p. 398-405.

[30] Liang G.-Z., Mei H., Zhou P., Zhou Y., Li Z.-L., *Study on quantitative structure-activity relationship by 3D holographic vector of atomic interaction field*, Acta Phys-Chim Sin 22(3), 2006, p. 388-390.

[31] Rosner B., *Fundamentals of Biostatistics*, 4th Edition, Duxbury Press, Belmont, California, USA, 1995.

[32] Katritzky A. R., Kuanar M., Slavov S., Dobchev D.A., Fara D. C., Karelson M., Acree Jr. W. E., Solov'ev V. P., Varnek A., *Correlation of blood-brain penetration using structural descriptors*, Bioorg Med Chem, 14(14), 2006, p. 4888-4917.

[33] Wang Y., Zhao C., Ma W., Liu H., Wang T., Jiang G., *Quantitative structure-activity relationship for prediction of the toxicity of polybrominated diphenyl ether (PBDE) congeners*, Chemosphere 64(4), 2006, p. 515-524.

[34] Roy D. R., Parthasarathi R., Subramanian V., Chattaraj P. K., *An electrophilicity based analysis of toxicity of aromatic compounds towards Tetrahymena pyriformis*, QSAR Comb Sci 25(2), 2006, p 114-122.

[35] Srivastava H. K., Pasha F. A., Singh P. P., *Atomic softness-based QSAR study of testosterone*, Int J Quantum Chem 103(3), 2005, p. 237-245.

[36] Xue C. X., Zhang R. S., Liu H. X., Yao X. J., Hu M. C., Hu Z. D., Fan B. T., *QSAR models for the prediction of binding affinities to human serum albumin using the heuristic method and a support vector machine*, J Chem Inf Comput Sci 44(5), 2004, p. 1693-1700.

[37] Jäntschi L., *Molecular Descriptors Family on Structure Activity Relationships 1. Review of the Methodology*, Leonardo Electronic Journal of Practices and Technologies 6, 2005, p. 76-98.

ET36/2005 – Et. Finală/2006 – Lucrare in extenso

toxicity [³⁸, ³⁹], mutagenicity [38], antioxidant efficacy [⁴⁰], antituberculous activity [⁴¹], antimalarial activity [⁴²], antiallergic activity [⁴³], anti-HIV-1 potencies [⁴⁴], inhibition activity on carbonic anhydrase II [⁴⁵] and IV [⁴⁶].

Several correlation coefficients based on different statistical hypothesis are known and most frequently used today: Pearson correlation coefficient, Spearman rank correlation

[38] Jäntschi L., Bolboacă S., *Molecular Descriptors Family on QSAR Modeling of Quinoline-based Compounds Biological Activities*, The 10th Electronic Computational Chemistry Conference 2005; http://bluehawk.monmouth.edu/~rtopper/eccc10_absbook.pdf as on 13 May 2006.

[39] Bobloacă S.D., Jäntschi L., *Modeling of Structure-Toxicity Relationship of Alkyl Metal Compounds by Integration of Complex Structural Information*, *Terapeutics, Pharmacology and Clinical Toxicology* X(1), 2006, p. 110-114.

[40] Bolboacă S., Filip C., Țigan Ș., Jäntschi L., *Antioxidant Efficacy of 3-Indolyl Derivates by Complex Information Integration*, *Clujul Medical* LXXIX(2), 2006, p. 204-209.

[41] Bolboacă S., Jäntschi L., *Molecular Descriptors Family on Structure Activity Relationships 3. Antituberculous Activity of some Polyhydroxyxanthenes*, *Leonardo Journal of Sciences* 7, 2005, p. 58-64.

[42] Jäntschi L., Bolboacă S., *Molecular Descriptors Family on Structure Activity Relationships 5. Antimalarial Activity of 2,4-Diamino-6-Quinazoline Sulfonamide Derivates*, *Leonardo Journal of Sciences* 8, 2006, p. 77-88.

[43] Jäntschi L., Bolboacă S., *Antiallergic Activity of Substituted Benzamides: Characterization, Estimation and Prediction*, *Clujul Medical* LXXIX, 2006, In press.

[44] Bolboacă S., Țigan Ș., Jäntschi L., *Molecular Descriptors Family on Structure-Activity Relationships on anti-HIV-1 Potencies of HEPTA and TIBO Derivatives*, In: Reichert A., Mihalaș G., Stoicu-Tivadar L., Schulz Ș., Engelbrech R. (Eds.), *Proceedings of the European Federation for Medical Informatics Special Topic Conference*, p. 222-226, 2006.

[45] Jäntschi L., Ungureșan M. L., Bolboacă S.D., *Integration of Complex Structural Information in Modeling of Inhibition Activity on Carbonic Anhydrase II of Substituted Disulfonamides*, *Applied Medical Informatics* 17(3,4), 2005, p. 12-21.

[46] Jäntschi L., Bolboacă S., *Modelling the Inhibitory Activity on Carbonic Anhydrase IV of Substituted Thiadiazole- and Thiadiazoline- Disulfonamides: Integration of Structure Information*, *Electronic Journal of Biomedicine*, 2006, In press.

ET36/2005 – Et. Finală/2006 – Lucrare in extenso

coefficient and Spearman semi-quantitative correlation coefficient, Kendall tau-a, -b and -c correlation coefficients, Gamma correlation coefficient [31].

Starting from the results obtained by applying of MDF-SAR methodology on a sample of sixty-seven compounds and from the assumption that the measured activity (compounds' inhibitory activity) of a biologically active compounds is a semi-quantitative outcome (can be related with the type of equipment used, the researchers, the chemical used), the abilities of Pearson, Spearman, Kendall's, and Gamma correlation coefficients in analysis of estimated toxicity were studied.

Multi-varied MDF-SAR model of pyrimidine derivatives

A sample of sixty-seven pyrimidine derivatives with inhibitory activity on E. coli dihydrofolate reductase (DHFR) was studied by the use of MDF-SAR methodology.

The set of pyrimidine derivatives (2,4-Diamino-5-(substituted-benzyl)-pyrimidine derivatives) with inhibitory activity on E. coli dihydrofolate reductase (DHFR) was previously studied by Ting-Lan Chiu & Sung-Sau So by the use of neural network approach [47].

By applying the MDF-SAR methodology on the sample of sixty-seven pyrimidine derivatives, a multi-varied model with four descriptors revealed to have good performances in prediction and estimation of inhibitory activity.

The multi-varied MDF-SAR model with four descriptors had the following equation:

$$Y_{\text{est}} = 3.78 + 1.62 \cdot iImrKHt + 2.37 \cdot liMDWHg + 6.40 \cdot IsDrJQt - 8.52 \cdot 10^{-2} \cdot LSPmEQg$$

Analyzing the MDF-SAR model with four descriptors it could be said that inhibitory activity considers compounds geometry (**g**) and topology (**t**), being related with the number of directly bonded hydrogens (**H**) of compounds and with the partial charge (**Q**) as atomic properties.

Statistical characteristics of the MDF-SAR model with four descriptors are in table 1 and 2.

Table 1. Statistical characteristics of the multi-varied MDF-SAR model with four descriptors

Characteristic (notation)	Value
---------------------------	-------

[47] Chiu T.L., So S. S., *Development of neural network QSPR models for Hansch substituent constants. 2. Applications in QSAR studies of HIV-1 reverse transcriptase and dihydrofolate reductase inhibitors*, J Chem Inf Comput Sci 44(1), 2004, p. 154-160.

Number of variable (v)	4
Correlation coefficient (r)	0.9517
95% Confidence Intervals for r (95% CI _r)	[0.9223, 0.9701]
Squared correlation coefficient (r ²)	0.9058
Adjusted squared correlation coefficient (r ² _{adj})	0.8997
Standard error of estimated (S _{est})	0.1919
Fisher parameter (F _{est})	149*
Cross-validation leave-one-out (loo) score (r ² _{cv-loo})	0.8932
Fisher parameter for loo analysis (F _{pred})	130*
Standard error for leave-one-out analysis (S _{loo})	0.2044
Model stability (r ² - r ² _{cv(loo)})	0.0126
r ² (iImrKHt, liMDWHg)	0.2020
r ² (iImrKHt, IsDrJQt)	0.0047
r ² (iImrKHt, LSPmEQg)	0.1482
r ² (liMDWHg, IsDrJQt)	0.0003
r ² (liMDWHg, LSPmEQg)	0.0212
r ² (IsDrJQt, LSPmEQg)	0.0664

*p < 0.001

Table 2. Statistics of the regression MDF-SAR model with four descriptors

	StdError	t Stat	95%CI _{coefficient}	r(Y _m ,desc)
Intercept	0.1999	18.92*	[3.38, 4.18]	n.a.
iImrKHt	0.0709	22.85*	[1.48, 1.76]	0.4803
liMDWHg	0.1500	15.81*	[2.07, 2.67]	0.0558
IsDrJQt	1.4779	4.33*	[3.45, 9.36]	0.0336
LSPmEQg	0.0182	-4.68*	[-0.12, -0.12]	0.0231

StdError = standard error; t Stat = Student tets parameter;
 95% CI_{coefficient} = 95% confidence interval associated with regression coefficients;
 Y_m = measured inhibitory activity; desc = molecular descriptor; * p < 0.001

Graphical representation of the measured versus estimated by MDF-SAR model with four descriptors inhibitory activity is in figure 1.

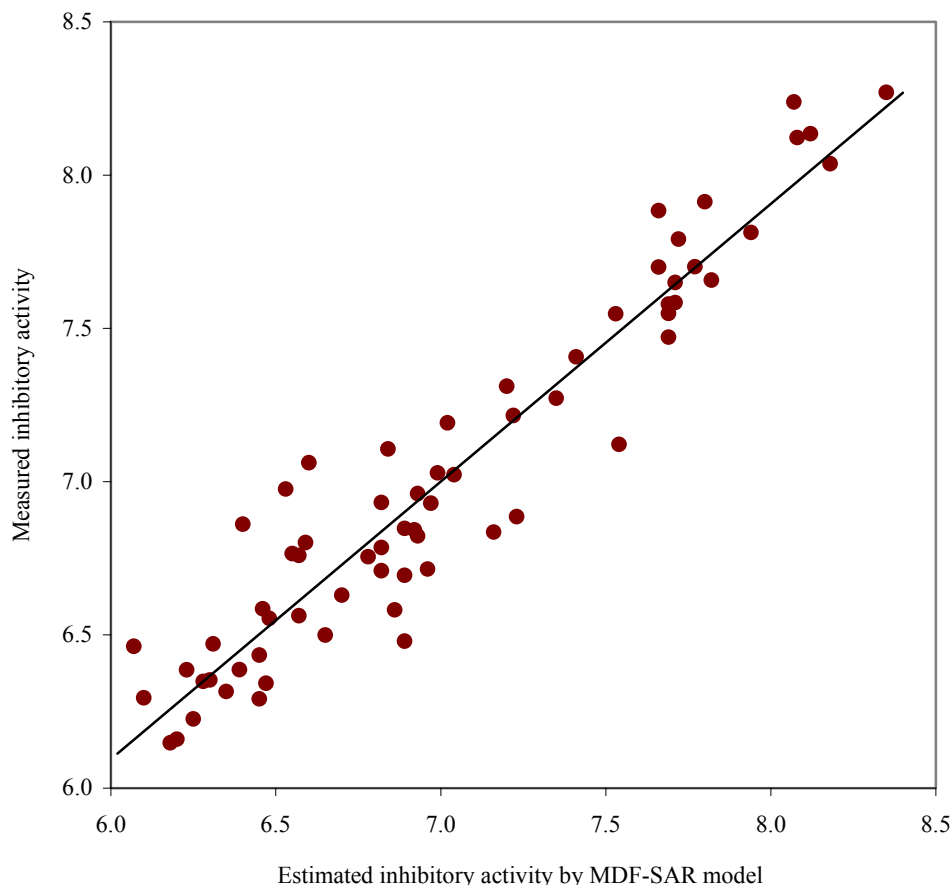


Figure 1. Plot of measured vs estimated by MDF-SAR inhibitory activity

Internal validation of the four-varied MDF SAR model with four descriptors was performed through splitting the whole set into training and test sets by applying of a randomization algorithm.

The coefficients for each model obtained in training sets, in conformity with the generic equation $Y_{est} = a_0 + a_1 \cdot iImrKHt + a_2 \cdot liMDWHg + a_3 \cdot IsDrJQt - a_4 \cdot 10^{-2} \cdot LSPmEQg$, the number of compounds in training (N_{tr}) and test (N_{ts}) sets, the correlation coefficient for training (r_{tr}) and test (r_{ts}) sets with associated 95% confidence intervals ($95\%CI_{tr}$ and $95\%CI_{ts}$), the Fisher parameter associated with training (F_{tr}) and test (F_{ts}) sets, and the Fisher's Z parameter of correlation coefficients comparison ($Z_{trtr-rts}$) are in table 3.

Table 3. Statistics results on training versus test sets

a_0	a_1	a_2	a_3	a_4	N_{tr}	r_{tr}	$95\%CI_{tr}$	F_{tr}	N_{ts}	r_{ts}	$95\%CI_{ts}$	F_{ts}	$Z_{trtr-rts}$
3.93	1.61	2.43	6.56	$-9.67 \cdot 10^{-2}$	35	0.949	[0.899, 0.974]	67*	32	0.958	[0.916, 0.980]	59*	0.418 [†]
3.98	1.57	2.45	6.55	$-7.52 \cdot 10^{-2}$	36	0.951	[0.905, 0.975]	73*	31	0.951	[0.899, 0.976]	61*	0.000 [†]
3.84	1.55	2.15	9.08	$-9.12 \cdot 10^{-2}$	37	0.944	[0.893, 0.908]	66*	30	0.949	[0.895, 0.976]	55*	0.206 [†]
3.94	1.59	2.42	6.10	$-8.18 \cdot 10^{-2}$	38	0.951	[0.907, 0.974]	78*	29	0.947	[0.890, 0.975]	50*	0.144 [†]
3.91	1.56	2.25	8.22	$-1.04 \cdot 10^{-1}$	39	0.963	[0.931, 0.981]	110*	28	0.937	[0.867, 0.971]	39*	1.069 [†]
4.18	1.51	2.44	6.06	$-7.22 \cdot 10^{-2}$	40	0.956	[0.917, 0.975]	92*	27	0.936	[0.863, 0.971]	35*	0.721 [†]
3.76	1.63	2.32	7.35	$-1.02 \cdot 10^{-1}$	41	0.963	[0.931, 0.980]	116*	26	0.935	[0.858, 0.971]	34*	1.104 [†]
3.97	1.58	2.39	5.11	$-9.36 \cdot 10^{-2}$	42	0.956	[0.919, 0.976]	99*	25	0.954	[0.896, 0.980]	34*	0.115 [†]

ET36/2005 – Et. Finală/2006 – Lucrare in extenso

3.64	1.64	2.30	7.00	$-8.15 \cdot 10^{-2}$	43	0.955	[0.917, 0.975]	98*	24	0.944	[0.873, 0.976]	37*	0.407 [†]
3.72	1.66	2.43	5.78	$-8.12 \cdot 10^{-2}$	44	0.938	[0.889, 0.966]	72*	23	0.964	[0.916, 0.985]	54*	1.030 [†]
3.59	1.64	2.25	4.94	$-9.98 \cdot 10^{-2}$	45	0.947	[0.904, 0.970]	86*	22	0.957	[0.898, 0.982]	37*	0.411 [†]
3.86	1.55	2.23	8.68	$-8.86 \cdot 10^{-2}$	46	0.940	[0.894, 0.967]	78*	21	0.983	[0.958, 0.993]	43*	2.290*
4.04	1.54	2.36	6.46	$-7.31 \cdot 10^{-2}$	47	0.949	[0.911, 0.972]	96*	20	0.963	[0.906, 0.985]	34*	0.538 [†]
3.63	1.63	2.24	4.27	$-8.93 \cdot 10^{-2}$	48	0.940	[0.895, 0.966]	82*	19	0.963	[0.904, 0.986]	44*	0.852 [†]
3.98	1.57	2.42	6.49	$-8.59 \cdot 10^{-2}$	49	0.946	[0.905, 0.969]	93*	18	0.960	[0.894, 0.985]	36*	0.535 [†]
3.77	1.61	2.32	6.37	$-8.46 \cdot 10^{-2}$	50	0.943	[0.902, 0.968]	91*	17	0.974	[0.927, 0.991]	52*	1.294 [†]
3.67	1.63	2.22	6.56	$-1.01 \cdot 10^{-1}$	51	0.954	[0.919, 0.973]	115*	16	0.950	[0.858, 0.983]	17*	0.126 [†]
3.81	1.61	2.39	6.87	$-7.70 \cdot 10^{-2}$	52	0.951	[0.916, 0.972]	112*	15	0.950	[0.853, 0.984]	22*	0.032 [†]
3.69	1.65	2.36	6.32	$-8.21 \cdot 10^{-2}$	53	0.953	[0.919, 0.972]	118*	14	0.956	[0.864, 0.986]	17*	0.128 [†]
3.97	1.56	2.40	6.16	$-7.51 \cdot 10^{-2}$	54	0.951	[0.916, 0.971]	115*	13	0.954	[0.851, 0.987]	17*	0.122 [†]

[†] p > 0.05; * p < 0.01

Definitions, Formulas, Interpretations, PHP functions, and Results

A number of add notations were used in the study, as follows:

- Pearson product-moment correlation coefficient (named after Karl Pearson (1857 - 1936), a major contributor to the early development of statistics):
 - r_{prs} = the *Pearson* correlation coefficient;
 - r_{Prs}^2 = the squared *Pearson* correlation coefficient;
 - $t_{Prs,df}$ = the Student test parameter, and its significance $p_{Prs,df}$ at a significance level of 5% (where df = the degree of freedom);
- Spearman's rank correlation coefficient (named after Charles Spearman (1863 - 1945), English psychologist known for his work in statistics - *factor analysis*, and *Spearman's rank correlation coefficient*):
 - r_{Spm} = the *Spearman* rank correlation coefficient
 - r_{Spm}^2 = the squared of *Spearman* rank correlation coefficient;
 - $t_{Prs,df}$ = the Student test parameter, and its significance $p_{Spm,df}$;
 - r_{sQ}^2 = the squared of *Spearman semi-Quantitative* correlation coefficient;
 - t_{sQ} = the Student test parameter, and its significance p_{sQ} ;
- Kendall's tau correlation coefficients (named after Maurice George Kendall (1907 - 1983), a prominent British statistician; published in monograph *Rank Correlation* in 1948);
 - $\tau_{Ken,a}$ = the *Kendall tau-a* correlation coefficient;
 - $\tau_{Ken,a}^2$ = the squared of *Kendall tau-a* correlation coefficient;
 - $Z_{Ken,ta}$ = the Z-test parameter of Kendall tau-a correlation coefficient, and its significance $p_{Ken,ta}$;
 - $\tau_{Ken,b}$ = the *Kendall tau-b* correlation coefficient;

ET36/2005 – Et. Finală/2006 – Lucrare in extenso

- $\tau_{\text{Ken,b}}^2$ = the squared of *Kendall tau-b* correlation coefficient;
- $Z_{\text{Ken,tb}}$ = the Z-test parameter of *Kendall tau-b*, and its significance $p_{\text{Ken,tb}}$;
- $\tau_{\text{Ken,c}}^2$ = the *Kendall tau-c* correlation coefficient;
- $\tau_{\text{Ken,c}}^2$ = the squared of *Kendall tau-c* correlation coefficient;
- $Z_{\text{Ken,tc}}$ = the Z-test parameter of *Kendall tau-c*, and its significance $p_{\text{Ken,tc}}$;
- Gamma correlation coefficient (also known as Goodman and Kruskal's gamma):
 - Γ = the *Gamma* correlation coefficient;
 - Γ^2 = the squared of *Gamma* correlation coefficient;
 - Z_{Γ} = the Z-test parameter of *Gamma* correlation coefficient, and its significance p_{Γ} .

A series of *.php programs which to facilitate the calculation and to display of above-described correlation coefficients and their statistics (Student-test and Z-test parameters and associated significances) were implemented and was use in order to reach the objective of study [48].

Pearson correlation coefficient

Definition: a measure the strength and direction of the linear relationship between two variables, describing the direction and degree to which one variable is linearly related to another.

Assumptions: both variable (variables Y_m and Y_{est}) are interval or ratio variables and are well approximated by a normal distribution, and their joint distribution is bivariate normal [49].

Formula

$$r_{\text{Prs}} = \frac{\sum (Y_{m-i} - \bar{Y}_m)(Y_{\text{est}-i} - \bar{Y}_{\text{est}})}{\sqrt{(\sum (Y_{m-i} - \bar{Y}_m)^2)(\sum (Y_{\text{est}-i} - \bar{Y}_{\text{est}})^2)}}$$

[48] ***Rank, ©2005, Virtual Library of Free Software, available at: http://vl.academicdirect.org/molecular_topology/mdf_findings/rank/

[49] ***Pearson's Correlation Coefficient [online], Available at: <http://www.texasoft.com/winkpear.html>

ET36/2005 – Et. Finală/2006 – Lucrare in extenso

where Y_{m-i} is the value of the measured inhibitory activity for compound i ($i = 1, 2, \dots, 67$)
 \bar{Y}_m is the average of the measured inhibitory activity, Y_{est-i} is the value of the estimated inhibitory activity for compound i , and \bar{Y}_{est} is the average of the estimated inhibitory activity.

Interpretation

The Pearson correlation coefficient can take values from -1 to +1. A value of +1 show that the variables are perfectly linear related by an increasing relationship, a value of -1 show that the variables are perfectly linear related by an decreasing relationship, and a value of 0 show that the variables are not linear related by each other. There is considered a strong correlation if the correlation coefficient is greater than 0.8 and a weak correlation if the correlation coefficient is less than 0.5.

The coefficient of determination (or r squared) gives information about the proportion of variation in the dependent variable which might be considered as being associated with the variation in the independent variable.

Related statistics

- The squared of Pearson correlation coefficient or Pearson coefficient of determination (r_{Prs}^2):
 - Describe the proportion of variance in Y_m that is related with linear variation of Y_{est} ;
 - Can take values from 0 to 1.

Statistical test

Student t-test was used to determine if the value of Pearson correlation coefficient is statistically significant, at a significance level of 5%.

The null hypothesis vs. the alternative hypothesis was:

$$H_0: r_{Prs} = 0 \text{ (there is no correlation between the variables)}$$

$$H_1: r_{Prs} < > 0 \text{ (variables are correlated)}$$

For a significance level equal with 5%, a p-value associated to $t_{Prs,df}$ less than 0.05 means that there is evidence to reject the null hypothesis in favor of the alternative hypothesis. In other words there is a statistically significant linear relationship between the variables.

ET36/2005 – Et. Finală/2006 – Lucrare in extenso
PHP implementation

In order to compute the statistics associated with Pearson correlation coefficient, three functions were implemented:

```
function coef_rk(&$y1,&$y2){
    $my1=m1($y1);
    $dy2=m2($y1,$y1)-$my1*$my1;
    $mx1=m1($y2);
    $mxy=m2($y2,$y1);
    $m2x=$mx1*$mx1;
    $mx2=m2($y2,$y2);
    $dx2=$mx2-$m2x;
    $r2=pow($mxy-$mx1*$my1,2)/($dx2*$dy2);
    return $r2;
}
function t_p($n,$k,$r){
    return $r*pow($n-$k-1,0.5)/pow(1-pow($r,2),0.5);
}
function p_t($t,$df){
    $p = $df/2;
    $x = 0.5+0.5*$t/pow(pow($t,2)+$df,0.5);
    $beta_gam = exp(-logBeta($p, $p) + $p * log($x) + $p * log(1.0 - $x) );
    return (2.0 * $beta_gam * betaFraction(1.0 - $x, $p, $p) / $p);
}
```

The statistics of Pearson correlation coefficients are computed as follows:

- Pearson correlation coefficient:

$$r_{pe} = coef_rk(\$cmp[0],\$cmp[1]);$$

where $\$cmp[0]$ is the measured inhibitory activity (Y_m), and $\$cmp[1]$ is the estimated by MDF-SAR model with four descriptor inhibitory activity (Y_{est}).

- t Student parameter:

$$t_{pe} = t_p(\$n,1,pow(\$r_{pe},0.5));$$

- Significance of t Student parameter

$$p_{pe} = p_t(t_{pe},\$n-2);$$

Results

$$r_{Prs}^2 = 0.9058$$

$$t_{Prs,1} = 24.99$$

$$p_{Prs,1} = 4.74 \cdot 10^{-33} \%$$

(1)

Definition

A non-parametric measure of correlation between variable which assess how well an arbitrary monotonic function could describe the relationship between two variables, without making any assumptions about the frequency distribution of the variables. Frequently the Greek letter ρ (rho) is use to abbreviate the Spearman correlation coefficient.

Spearman's rank correlation is satisfactory for testing the null hypothesis of no relationship, but is difficult to interpret as a measure of the strength of the relationship [50].

Assumptions

- Does not required any assumptions about the frequency distribution of the variables;
- Does not required the assumption that the relationship between variable is linear;
- Does not required the variable to be measured on interval or ration scale.

Formula

In order to compute the Spearman rank correlation coefficient, the two variables (Y_m , respectively Y_{est}) were converted to ranks (see table 4 for exemplification). For each measured and estimated inhibitory activity a rank was assigned ($RankY_m$ - for measure inhibitory activity, $RankY_{est}$ - for estimated by MDF-SAR model inhibitory activity) according with the position of value into a sort serried of values.

In assignment of rank process, the lowest value had the lowest rank. When there are two equal values for two different compounds (for measured and/or estimated inhibitory activity), the associated rank had equal values and was calculated as means of corresponding ranks. For example, the compounds abbreviated as c_52 and c_59 have the same measured inhibitory activity (6.45, see table 4). The rank associated with these values is equal with 13.5 (is the average between the rank for c_52 - 13 and the rank of c_59 - 14).

Table 4. Compounds abbreviation, measured and estimated activity and associated ranks

Abb.	Y_m	$RankY_m$		Y_{est}	$RankY_{est}$	Abb.	Y_m	$RankY_m$		Y_{est}	$RankY_{est}$
c_64	6.07	1	0	6.4626	13	c_32	6.92	35	0	6.8423	32
c_65	6.10	2	0	6.2948	5	c_66	6.93	36.5	5	6.8225	30
c_67	6.18	3	0	6.1479	1	c_36	6.93	36.5		6.9609	38

[50] Methods based on rank order. In: Bland M., *An Introduction to Medical Statistics*, Oxford University Press; Oxford, New York, Tokyo, p. 205-225, 1995.

ET36/2005 – Et. Finală/2006 – Lucrare in extenso

c_54	6.20	4	0	6.1595	2	c_40	6.96	38	0	6.7150	24
c_37	6.23	5	0	6.3859	10	c_17	6.97	39	0	6.9298	36
c_48	6.25	6	0	6.2254	3	c_45	6.99	40	0	7.0283	41
c_31	6.28	7	0	6.3483	8	c_41	7.02	41	0	7.1919	45
c_49	6.30	8	0	6.3528	9	c_15	7.04	42	0	7.0225	40
c_10	6.31	9	0	6.4703	14	c_28	7.16	43	0	6.8355	31
c_56	6.35	10	0	6.3149	6	c_09	7.20	44	0	7.3115	48
c_47	6.39	11	0	6.3866	11	c_18	7.22	45	0	7.2156	46
c_53	6.40	12	0	6.8614	34	c_43	7.23	46	0	6.8855	35
c_52	6.45	13.5	1	6.2913	4	c_29	7.35	47	0	7.2724	47
c_59	6.45	13.5		6.4336	12	c_14	7.41	48	0	7.4072	49
c_16	6.46	15	0	6.5851	20	c_24	7.53	49	0	7.5476	51
c_34	6.47	16	0	6.3422	7	c_22	7.54	50	0	7.1218	44
c_58	6.48	17	0	6.5536	17	c_26	7.66	51.5	6	7.7002	57
c_35	6.53	18	0	6.9755	39	c_08	7.66	51.5		7.8841	61
c_42	6.55	19	0	6.7654	27	c_27	7.69	54	7	7.4715	50
c_30	6.57	20.5	2	6.5625	18	c_13	7.69	54		7.5489	52
c_61	6.57	20.5		6.7594	26	c_12	7.69	54		7.5793	53
c_33	6.59	22	0	6.8010	29	c_04	7.71	56.5	8	7.5841	54
c_51	6.60	23	0	7.0616	42	c_11	7.71	56.5		7.6497	55
c_39	6.65	24	0	6.4993	16	c_19	7.72	58	0	7.7915	59
c_38	6.70	25	0	6.6297	21	c_23	7.77	59	0	7.7014	58
c_57	6.78	26	0	6.7552	25	c_25	7.80	60	0	7.9130	62
c_60	6.82	28	3	6.7091	23	c_01	7.82	61	0	7.6576	56
c_44	6.82	28		6.7847	28	c_21	7.94	62	0	7.8130	60
c_55	6.82	28		6.9318	37	c_06	8.07	63	0	8.2391	66
c_20	6.84	30	0	7.1067	43	c_03	8.08	64	0	8.1224	64
c_46	6.86	31	0	6.5813	19	c_07	8.12	65	0	8.1353	65
c_50	6.89	33	4	6.4794	15	c_05	8.18	66	0	8.0372	63
c_62	6.89	33		6.6942	22	c_02	8.35	67	0	8.2702	67
c_63	6.89	33		6.8475	33						

The method of rank assignment for more than two equal values of measured and/or estimated inhibitory activity is the same as for two equal values. If there are an odd number of compounds which have the same measured value (see compounds c_60, c_44, and c_55 from table 2) then the rank will be an integer $((27+28+29)/3 = 28$, see the rank for c_60, c_44, and c_55).

In studied example, there are equal values for measured activity: five situations of two equal values (c_52-c_59, c_30-c_61, c_66-c_36, c_26-c_08, and c_04-c_11), and three situations of three equal values (c_60-c_44-c_55, c_50-c_62-c_63, and c_27-c_13-c_12).

By conversion of the measured and estimated inhibitory activity to ranks, the distribution of ranks does not depend on the distribution of measured, respectively estimated inhibitory activity.

The formula for calculation of the Spearman rank correlation coefficient is:

$$r_{\text{Spm}} = \frac{\sum (R_{Y_{m-i}} - \bar{R}_{Y_m})(R_{Y_{\text{est-i}}} - \bar{R}_{Y_{\text{est}}})}{\sqrt{(\sum (R_{Y_{m-i}} - \bar{R}_{Y_m})^2)(\sum (R_{Y_{\text{est-i}}} - \bar{R}_{Y_{\text{est}}})^2)}}$$

where $R_{Y_{m-i}}$ is the rank of the measured inhibitory activity for compound i , $\bar{R}_{Y_{m-i}}$ is the average of the measured inhibitory activity, $R_{Y_{\text{est-i}}}$ is the rank of the estimated by MDF-SAR inhibitory activity for compound i , and $\bar{R}_{Y_{\text{est-i}}}$ is the average of the estimated inhibitory activity.

The simple formula for r_{Spm} is based on the difference between each pairs of ranks:

$$r_{\text{Spm}} = 1 - \frac{6 \sum D^2}{n(n^2 - 1)}$$

where D is the differences between each pair of ranks (e.g. $D = R_{Y_{m-1}} - R_{Y_{\text{est-1}}}$) and n is the volume of the sample.

The formula of the *Spearman semi-quantitative method* is:

$$r_{\text{sQ}} = \sqrt{\frac{\sum (Y_{m-i} - \bar{Y}_m)(Y_{\text{est-i}} - \bar{Y}_{\text{est}})}{\sqrt{(\sum (Y_{m-i} - \bar{Y}_m)^2)(\sum (Y_{\text{est-i}} - \bar{Y}_{\text{est}})^2)}} \cdot \frac{\sum (R_{Y_{m-i}} - \bar{R}_{Y_m})(R_{Y_{\text{est-i}}} - \bar{R}_{Y_{\text{est}}})}{\sqrt{(\sum (R_{Y_{m-i}} - \bar{R}_{Y_m})^2)(\sum (R_{Y_{\text{est-i}}} - \bar{R}_{Y_{\text{est}}})^2)}}$$

Interpretation

- Identical with Pearson correlation coefficient.

Related statistics

- r_{Spm}^2 = the squared of *Spearman* rank correlation coefficient;
- r_{sQ}^2 = the squared of semi-quantitative correlation coefficient.

Statistical significance

- Compute by the use of a permutation test (a statistical test in which the reference distribution is obtained by permuting the observed data points across all possible outcomes, given a set of conditions consistent with the null hypothesis);
- Comparing the observed r_{Spm} with published tables for different levels of significance (eg. 0.05, 0.01...). It is a simple solution when the researchers want to know the significance within a certain range or less than a certain value;
- Tested by applying the Student t-test (for sample sizes > 20): the method used in this study.

ET36/2005 – Et. Finală/2006 – Lucrare in extenso

The null hypothesis vs. the alternative hypothesis for Spearman rank correlation coefficient was:

$$H_0: r_{\text{Spm}} = 0 \text{ (there is no correlation between the ranked pairs)}$$

$$H_1: r_{\text{Spm}} < > 0 \text{ (ranked pairs are correlated)}$$

The null hypothesis vs. the alternative hypothesis for semi-quantitative correlation coefficient was:

$$H_0: r_{\text{sQ}} = 0 \text{ (there is no correlation between the ranked pairs)}$$

$$H_1: r_{\text{sQ}} < > 0 \text{ (ranked pairs are correlated)}$$

PHP implementation

The formulas for Spearman and respectively semi-quantitative correlation coefficients used two defined above functions (t_p and respectively p_t). The Spearman rank correlation coefficient used the *coef_rk* function defined as:

```
function coef_rk(&$y1,&$y2){
    $my1=m1($y1);
    $dy2=m2($y1,$y1)-$my1*$my1;
    $mx1=m1($y2);
    $mxy=m2($y2,$y1);
    $m2x=$mx1*$mx1;
    $mx2=m2($y2,$y2);
    $dx2=$mx2-$m2x;
    $r2=pow($mxy-$mx1*$my1,2)/($dx2*$dy2);
    return $r2;
}
```

where

```
function m1(&$v){
    $rez=0;
    $n=count($v);
    for($i=1;$i<$n;$i++)
        $rez+=$v[$i];
    return $rez/($n-1);
}
function m2(&$v,&$u){
    $rez=0;
    $n=count($v);
    for($i=1;$i<$n;$i++)
        $rez+=$v[$i]*$u[$i];
    return $rez/($n-1);
}
```

ET36/2005 – Et. Finală/2006 – Lucrare in extenso
Spearman correlation coefficient

The statistics of Spearman rank correlation coefficients are computed as follows:

- Spearman correlation coefficient:

$$r_{sp} = coef_rk(\$poz[0],\$poz[1]);$$

where $\$poz[0]$ is the position on sort series of measured inhibitory activity, and $\$poz[1]$ is the position on sort series of estimated inhibitory activity by MDF-SAR model with four descriptor.

- t Student parameter:

$$t_{sp} = t_p(\$n,1,pow(r_{sp},0.5));$$

- Significance of t Student parameter

$$p_{sp} = p_t(t_{sp},\$n-2);$$

Semi-quantitative correlation coefficient

The statistics of semi-quantitative correlation coefficients are computed as follows:

- Semi-quantitative correlation coefficient:

$$r_{sq} = pow(r_{pe} * r_{sp}, 0.5);$$

- t Student parameter:

$$t_{sq} = t_p(\$n,1,pow(r_{sq},0.5));$$

- Significance of t Student parameter

$$p_{sq} = p_t(t_{sq},\$n-2);$$

Results

$$\begin{aligned} r_{Spr}^2 &= 0.8606 \\ t_{Spm,1} &= 20.03 \end{aligned} \tag{2}$$

$$\begin{aligned} p_{Spm,1} &= 1.62 \cdot 10^{-29} \\ r_{sQ}^2 &= 0.8829 \\ t_{sQ} &= 22.14 \\ p_{sQ} &= 5.57 \cdot 10^{-32} \end{aligned} \tag{3}$$

ET36/2005 – Et. Finală/2006 – Lucrare in extenso
Kendall's rank correlation coefficients

Definition

Kendall-tau is a non-parametric correlation coefficient that can be used to assess and test correlations between non-interval scaled ordinal variables. Frequently the Greek letter τ (tau), is use to abbreviate the Kendall tau correlation coefficient.

The Kendall tau correlation coefficient is considered to be equivalent to the Spearman rank correlation coefficient. While Spearman rank correlation coefficient is like the Pearson correlation coefficient but computed from ranks, the Kendall tau correlation rather represents a probability.

There are three Kendall's tau correlation coefficient known as tau-a, tau-b, and tau-c.

Formula

Let (Y_{m-i}, Y_{est-i}) and (Y_{m-j}, Y_{est-j}) be the pair of measured and estimated inhibitory activity. If $Y_{m-j} - Y_{m-i}$ and $Y_{est-j} - Y_{est-i}$, where $i < j$ have the same sign the pair is *concordant*, if have opposite signs the pair is *discordant*.

In a sample of n observations it can be found $n(n-1)/2$ pairs corresponding to choices $1 \leq i < j \leq n$.

The formulas of Kendall's tau correlation coefficients are as follows:

- Kendall tau-a correlation coefficient ($\tau_{Ken,a}$):

$$\tau_{Ken,a} = (C-D)/[n(n-1)/2]$$

- Kendall tau-b correlation coefficient ($\tau_{Ken,b}$):

$$\tau_{Ken,b} = (C-D)/\sqrt{[(n(n-1)/2-t)(n(n-1)/2-u)]}$$

where t is the number of tied Y_m values and u is the number of tied Y_{est} values.

- Kendall tau-c correlation coefficient ($\tau_{Ken,c}$):

$$\tau_{Ken,c} = 2(C-D)/n^2$$

Interpretation

- If the agreement between the two rankings is perfect and the two rankings are the same, the coefficient has value 1.
- If the disagreement between the two rankings is perfect and one ranking is the reverse of the other, the coefficient has value -1.

ET36/2005 – Et. Finală/2006 – Lucrare in extenso

- For all other arrangements the value lies between -1 and 1, and increasing values imply increasing agreement between the rankings.
- If the rankings are independent, the coefficient has value 0.

Related statistics

- $\tau_{\text{Ken,a}}^2$ = the squared of Kendall tau-a correlation coefficient;
- $\tau_{\text{Ken,b}}^2$ = the squared of Kendall tau-b correlation coefficient;
- $\tau_{\text{Ken,c}}^2$ = the squared of Kendall tau-c correlation coefficient.

Statistical significance

Statistical significance of the Kendall's tau correlation coefficient is tested by the Z-test, at a significance level of 5%. The null hypothesis vs. the alternative hypothesis for Kendall's tau correlation coefficients was:

- Kendall tau-a correlation coefficient:

$H_0: \tau_{\text{Ken,a}} = 0$ (there is no correlation between the two variables)

$H_1: \tau_{\text{Ken,a}} < > 0$ (the two variables are correlated)

- Kendall tau-b correlation coefficient:

$H_0: \tau_{\text{Ken,b}} = 0$ (there is no correlation between the two variables)

$H_1: \tau_{\text{Ken,b}} < > 0$ (the two variables are correlated)

- Kendall tau-c correlation coefficient:

$H_0: \tau_{\text{Ken,c}} = 0$ (there is no correlation between the two variables)

$H_1: \tau_{\text{Ken,c}} < > 0$ (the two variables are correlated)

PHP implementation

Kendall function was implemented in order to calculate the Kendall's tau correlation coefficients:

```
function Kendall(&$cmp){
    $n = count($cmp[0]);
    $pz = 0;
    if(!is_numeric($cmp[0][0])) $pz = 1;
    $C = 0;
    $D = 0;
    $E = 0;
    for($i=$pz;$i<$n-1;$i++){
        for($j=$i+1;$j<$n;$j++){
            $sgx = 0;
            $sgy = 0;
```

ET36/2005 – Et. Finală/2006 – *Lucrare in extenso*

```

if($cmp[0][$i]>$cmp[0][$j]) $sgx = 1;
if($cmp[0][$i]<$cmp[0][$j]) $sgx = -1;
if($cmp[1][$i]>$cmp[1][$j]) $sgy = 1;
if($cmp[1][$i]<$cmp[1][$j]) $sgy = -1;
if($sgx*$sgy>0) $C++;
if($sgx*$sgy<0) $D++;
if($sgx*$sgy==0) $E++;
if($sgx==0)$tied_x[$i][]=$j;
if($sgy==0)$tied_y[$i][]=$j;
}
}
$st1 = 0;      $su1 = 0;
$svt = 0;      $svu = 0;
$sv2t = 0;     $sv2u = 0;
if(isset($tied_x))
if(is_array($tied_x)){
    foreach($tied_x as $vx){
        $snt = count($vx)+1;
        $st1 += $snt*($snt-1);
        $svt += $snt*($snt-1)*(2*$snt+5);
        $sv2t += $snt*($snt-1)*($snt-2);
    }
}
if(isset($tied_y))
if(is_array($tied_y)){
    foreach($tied_y as $vy){
        $snu = count($vy)+1;
        $su1 += $snu*($snu-1);
        $svu += $snu*($snu-1)*(2*$snu+5);
        $sv2u += $snu*($snu-1)*($snu-2);
    }
}
}
$sv1 = $st1*$su1;
$st1 /= 2;
$su1 /= 2;
$sv2 = $sv2t*$sv2u;
$S = $C - $D;
$N = $n - $pz;
$scn2 = $N*($N-1)/2;
$stau_a2 = pow($S,2)/pow($scn2,2);
$sv_tau_a = $scn2*(2*$N+5)/9;
$sz_tau_a = $S/pow($sv_tau_a,0.5);
$T = ($scn2-$st1)*($scn2-$su1);
$stau_b2 = pow($S,2)/$T;
$svT0 = $sv_tau_a - ($svt + $svu)/18;
$svT1 = $sv1/(4*$scn2);
$svT2 = $sv2/(18*$scn2*($N-2));
$sv_tau_b = pow($svT0 + $svT1 + $svT2 , 0.5);
$sz_tau_b = $S/$sv_tau_b;
$gamma = pow(($C - $D)/($C + $D),2);
$sv_gamma = (2*$N+5)/9.0/$scn2;

```

ET36/2005 – Et. Finală/2006 – Lucrare in extenso

```
$z_gamma = $gamma/pow($v_gamma,0.5);  
$tau_c2 = 4*pow($S,2)/pow($n,4);  
$z_tau_c = $z_tau_b*($n-1)/$n;  
return array( $tau_a2, $z_tau_a, $tau_b2, $z_tau_b,  
$tau_c2, $z_tau_c, $gamma, $z_gamma );  
}
```

where C is the number of concordant pairs ($C = (<, <) \text{ or } (>, >)$), D is the number of discordant pairs ($D = (<, >) \text{ or } (>, <)$), and E is the number of equal pairs ($E = (=, .) \text{ or } (., =)$).

Results

- Kendall's τ_a correlation coefficient and associated statistics:

$$\begin{aligned}\tau_{\text{Ken},a}^2 &= 0.6129 \\ Z_{\text{Ken},\tau_a} &= 9.37 \\ p_{\text{Ken},\tau_a} &= 7.44 \cdot 10^{-21}\end{aligned}\tag{4}$$

- Kendall's τ_b correlation coefficient and associated statistics:

$$\begin{aligned}\tau_{\text{Ken},b}^2 &= 0.6177 \\ Z_{\text{Ken},\tau_b} &= 9.37 \\ p_{\text{Ken},\tau_b} &= 7.26 \cdot 10^{-21}\end{aligned}\tag{5}$$

- Kendall's τ_c correlation coefficient and associated statistics:

$$\begin{aligned}\tau_{\text{Ken},c}^2 &= 0.5948 \\ Z_{\text{Ken},\tau_c} &= 9.23 \\ p_{\text{Ken},\tau_c} &= 2.70 \cdot 10^{-20}\end{aligned}\tag{6}$$

Gamma correlation coefficient

Definition

The Gamma correlation coefficient (Γ , gamma) is a measure of association between variables that comparing with Kendall's tau correlation coefficients is more resistant to tied data [51], being preferable to Spearman rank or Kendall tau when data contain many tied observations [52].

[51] Goodman L. A., Kruskal W.H., *Measures of association for cross-classifications III: Approximate sampling theory*, J. Amer. Statistical Assoc. 58, 1963, p. 310-364.

[52] Siegel S., Castellan N. J., *Nonparametric Statistics for the Behavioural Sciences*, 2nd Edition, McGraw-Hill, 1988.

Formula

The formula for Gamma correlation coefficient is:

$$\Gamma = (C-D)/(C+D)$$

where the significance of C and D were described above.

Interpretation

- In the same manner as the Kendall tau correlation coefficient.

Related statistics

- Γ^2 = the squared of Gamma correlation coefficient.

Statistical significance

Statistical significance of Gamma correlation coefficient was tested by the Z-test, at a significance level of 5%. The null hypothesis vs. the alternative hypothesis for Gamma correlation coefficients was:

$H_0: \Gamma = 0$ (there is no correlation between the two variables)

$H_1: \Gamma < > 0$ (the two variables are correlated).

PHP implementation

The function which computes the Gamma correlation coefficient was presented at Kendall's tau correlation coefficient, in PHP implementation section.

Results

$$\Gamma^2 = 0.6208$$

$$Z_{\Gamma} = 7.43$$

$$p_{\Gamma} = 1.11 \cdot 10^{-13}$$

(7)

Conclusions

All seven computational methods used to evaluate the correlation between measured and estimated by MDF-SAR model inhibitory activity are statistically significant (p-value always less than 0.0001, correlation coefficients always greater than 0.5).

More research on other classes of biologic active compounds may reveal whether it is appropriate to analyze the MDF-SAR models using the Pearson correlation coefficient or other correlation coefficients (Spearman rank, Kendall's tau, or Gamma correlation coefficient).

Molecular Descriptors Family on Structure Activity Relationships

6. Octanol-Water Partition Coefficient of Polychlorinated Biphenyls

Abstract

Octanol-water partition coefficient of two hundred and six polychlorinated biphenyls was model by the use of an original method based on complex information obtained from compounds structure. The regression analysis shows that best results are obtained in four-varied model ($r^2 = 0.9168$). The prediction ability of the model was studied through leave-one-out analysis ($r^2_{cv(100)} = 0.9093$) and in training and test sets analysis. Modeling the octanol-water partition coefficient of polychlorinated biphenyls by integration of complex structural information provide a stable and performing four-varied model, allowing us to make remarks about relationship between structure of polychlorinated biphenyls and associated octanol-water partition coefficients.

Keywords

PolyChlorinated Biphenyls (PCBs), Molecular Descriptors Family (MDF), Structure-Property Relationships (SAR), Octanol-water partition coefficient

Background

Polychlorinated biphenyls (PCBs), stable organic industrials chemicals widely used as insulating fluids, hydraulic and lubricating fluids, heat exchanger fluids and as additives in adhesive inks and paints [53] are persistent in the environment [54] as well as in the living tissue [55].

[53] George C. J., Bennett G. F., Simoneaux D., George W. J., *Polychlorinated biphenyls. A toxicological review*, Journal of Hazardous Materials, 18(2), p. 113-144, 1988.

[54] Borja J., Taleon D. M., Auresenia J., Gallardo S., *Polychlorinated biphenyls and their biodegradation*, Process Biochemistry, 40(6), p. 1999-2013, 2005.

[55] Hertz-Picciotto I., Charles M. J., James R. A. , Keller J. A., Willman E., Teplin S., *In utero polychlorinated biphenyl exposures in relation to fetal and early childhood growth*, Epidemiology, 16(5), p. 648-656, 2005.

ET36/2005 – Et. Finală/2006 – Lucrare in extenso

Quantitative structure-property relationships of PCBs were previously studied taking into consideration octanol-water partition coefficients and soil-water partition coefficients [⁵⁶] and/or other physicochemical properties [⁵⁷].

Based on the complex information offered by the structure of polychlorinated biphenyls congeners, octanol-water partition coefficients expressed as $\log K_{ow}$ was modeled by applying of an original methodology. Thus, the aim of the paper is to present the performances of the original methodology in estimation and prediction of octanol-water partition coefficients of polychlorinated biphenyls.

Materials and Methods

A set of two-hundred and six polychlorinated biphenyls congeners with measured octanol-water partition coefficients were included into analysis. The values for the octanol-water partition coefficients were taken from a previous reported study [⁵⁸]. There were included ten PCBs congener group: mono-, di-, tri-, tetra-, penta-, hexa-, hepta-, octa-, nona-, decachlorobiphenyl. Table 1 contains the PCBs number, the structure (chlorine-filled) and associated octanol-water partition coefficients (expressed as $\log K_{ow}$).

The original methodology is based on molecular descriptors family computed based on the structure of the PCBs. The steps used to model the activity of interest were presented in details on [⁵⁹] and were:

- Step 1: Sketch of the three-dimensional structure of polychlorinated biphenyls congeners;

[56] Hansen B. G., Paya-Perez A. B., Rahman M., Larsen B. R., *QSARs for $K(ow)$ and $K(oc)$ of PCB congeners: A critical examination of data, assumptions and statistical approaches*, *Chemosphere*, 39(13), p. 2209-2228, 1999.

[57] Zou J.-W., Jiang Y.-J., Hu G.-X., Zeng M., Zhuang S.-L., Yu Q.-S., *QSPR/QSAR studies on the physicochemical properties and biological activities of polychlorinated biphenyls*, *Acta Physico-Chimica Sinica*, 21(3), p. 267-272, 2005.

[58] Eisler R., Belisle A. A., *Planar PCB Hazards to Fish, Wildlife, and Invertebrates: A Synoptic Review*, *Contaminant Hazard Reviews*, p. 1-96, 1996.

[59] Jäntschi L., *Molecular Descriptors Family on Structure Activity Relationships 1. The review of Methodology*, *Leonardo Electronic Journal of Practices and Technologies*, AcademicDirect, 6, p. 76-98, 2005.

ET36/2005 – Et. Finală/2006 – *Lucrare in extenso*

- Step 2: Create the file with the measured octanol-water partition coefficients of the polychlorinated biphenyls congeners;
- Step 3: Generating, computing and filtering the members of molecular descriptors family for polychlorinated biphenyls congeners;
- Step 4: Finding and identifying the SAR models for polychlorinated biphenyls congeners;
- Step 5: Validate the SAR model by a cross-validation analysis [60];
- Step 6: Analyze the selected SAR model.

Results and Discussions

Modeling of the octanol-water partition coefficients of the polychlorinated biphenyls congeners was run on mono-, bi-, and tetra-varied SARs. The model which obtained best performance was the four-varied model and is presented here. The equation of the four varied model is:

$$\hat{Y}_{\log K_{ow}} = 3.039 - 0.421 \cdot IIDDKGg + 0.044 \cdot IHDRKEg + 0.070 \cdot aHMmjQt - 37.502 \cdot aSMMjQg$$

The abbreviation associated with the studied PCBs congener (PBC no.), the measured octanol-water partition coefficients (express as $\log K_{ow}$), the values of the descriptors used and estimated octanol-water partition coefficients by the model ($\hat{Y}_{\log K_{ow}}$) and the absolute differences between estimated by the model and measured octanol-water partition coefficients ($|\hat{Y} - \log K_{ow}|$) are in table 1.

Table 1. Polychlorinated biphenyls abbreviation, $\log K_{ow}$, values for descriptors used by model, $\hat{Y}_{\log K_{ow}}$, and $|\hat{Y}_{\log K_{ow}} - \log K_{ow}|$

PCB no.	Structure (chlorine-filled)	$\log K_{ow}$	IIDDKGg	IHDRKEg	aHMmjQt	aSMMjQg	$\hat{Y}_{\log K_{ow}}$	$ \hat{Y} - \log K_{ow} $
1	2	4.6010	5.7503	91.1540	0.0244	$3.67 \cdot 10^{-5}$	4.6477	0.0467
2	3	4.4210	6.8329	100.870	0.0286	$5.60 \cdot 10^{-4}$	4.6022	0.1812
3	4	4.4010	7.1099	105.020	0.0303	$1.50 \cdot 10^{-4}$	4.6845	0.2835
4	2,2'	5.0230	5.6688	98.0270	0.0454	$2.17 \cdot 10^{-4}$	4.9804	0.0426
5	2,3'	5.0210	6.0092	104.370	0.1765	$4.61 \cdot 10^{-5}$	5.1330	0.1120
6	2,4	5.1500	7.0663	113.130	0.0205	$4.01 \cdot 10^{-5}$	5.0646	0.0854
7	2,4'	5.3010	7.2970	115.000	0.0079	$8.57 \cdot 10^{-5}$	5.0476	0.2534
8	2,5	5.1800	5.8788	102.720	0.1013	$4.82 \cdot 10^{-5}$	5.1096	0.0704
9	2,6	5.3110	5.3684	95.9580	0.1265	$4.87 \cdot 10^{-5}$	5.0274	0.2836
10	3,3'	5.3430	6.8183	115.510	0.0464	$2.30 \cdot 10^{-4}$	5.2688	0.0742
11	3,4	5.2950	7.2304	118.150	0.0067	$4.30 \cdot 10^{-5}$	5.2163	0.0787

[60] ***, *Leave-one-out Analysis*, © 2005, Virtual Library of Free Software, available at: http://vl.academicdirect.org/molecular_topology/mdf_findings/loo.

ET36/2005 – Et. Finală/2006 – Lucrare in extenso

12	3,5	5.4040	6.7261	115.520	0.0357	5.34·10 ⁻⁴	5.2959	0.1081
13	4,4'	5.3350	7.3646	124.560	0.0065	1.00·10 ⁻⁴	5.4409	0.1059
14	2,2',3	5.3110	6.5110	114.030	0.0668	2.44·10 ⁻⁴	5.3336	0.0226
15	2,2',4	5.7610	6.9032	119.960	0.0867	2.52·10 ⁻⁴	5.4317	0.3293
16	2,2',5	5.5510	6.8266	120.050	0.0579	2.66·10 ⁻⁴	5.4654	0.0856
17	2,2',6	5.4810	5.8973	106.490	0.0865	3.80·10 ⁻⁴	5.2550	0.2260
18	2,3,3'	5.5770	7.8504	128.320	0.7667	8.00·10 ⁻⁵	5.4564	0.1206
19	2,3,4	5.5170	7.4010	125.250	0.0235	5.28·10 ⁻⁵	5.4591	0.0579
20	2,3,4'	5.4210	8.1564	132.790	0.0508	7.75·10 ⁻⁵	5.4753	0.0543
21	2,3,5	5.5770	6.5446	122.510	0.7310	1.40·10 ⁻⁴	5.7444	0.1674
22	2,3,6	5.6710	6.2218	110.890	0.0325	1.01·10 ⁻⁴	5.3195	0.3515
23	2,3',4	5.6770	8.1210	134.330	0.0325	5.58·10 ⁻⁵	5.5578	0.1192
24	2,3',5	5.6670	7.3770	129.310	0.1674	8.11·10 ⁻⁵	5.6575	0.0095
25	2,3',6	5.4470	6.2046	113.400	0.0385	1.01·10 ⁻⁴	5.4381	0.0089
26	2,4,4'	5.6910	8.4470	139.210	0.0266	6.66·10 ⁻⁵	5.6355	0.0555
27	2,4,5	5.7430	7.1079	126.640	0.0245	5.28·10 ⁻⁵	5.6439	0.0991
28	2,4,6	5.5040	6.5149	114.430	0.0420	1.50·10 ⁻⁴	5.3515	0.1525
29	2,4',5	5.6770	7.5935	133.180	0.0270	7.78·10 ⁻⁵	5.7278	0.0508
30	2,4',6	5.7510	6.4853	116.740	0.0429	1.60·10 ⁻⁴	5.4657	0.2853
31	2',3,4	5.5720	7.7422	129.640	0.0889	7.82·10 ⁻⁵	5.5131	0.0589
32	2',3,5	5.6670	7.2479	126.270	0.0128	7.99·10 ⁻⁵	5.5668	0.1002
33	3,3',4	5.8270	8.6854	141.730	0.1163	3.69·10 ⁻⁴	5.6414	0.1856
34	3,3',5	4.1510	8.1095	137.500	0.3422	3.53·10 ⁻²	4.4021	0.2511
35	3,4,4'	4.9410	8.9668	146.520	0.3565	1.86·10 ⁻⁴	5.7583	0.8173
36	3,4,5	5.7670	7.0857	129.590	0.0297	1.89·10 ⁻³	5.7151	0.0519
37	3,4',5	5.8970	8.3823	142.120	0.0550	3.49·10 ⁻⁴	5.7827	0.1143
38	2,2',3,3'	5.5610	7.7406	134.940	0.1340	2.94·10 ⁻⁴	5.7430	0.1820
39	2,2',3,4	6.1110	7.8567	137.950	0.0258	2.88·10 ⁻⁴	5.8198	0.2912
40	2,2',3,4'	5.7670	8.0115	139.920	0.1247	2.83·10 ⁻⁴	5.8488	0.0818
41	2,2',3,5	5.7570	7.2531	133.770	0.0875	2.68·10 ⁻⁴	5.8942	0.1372
42	2,2',3,5'	5.8110	7.5381	135.540	0.0446	3.09·10 ⁻⁴	5.8479	0.0369
43	2,2',3,6	5.5370	6.5064	121.120	0.1066	3.47·10 ⁻⁴	5.6478	0.1108
44	2,2',3,6'	5.5370	6.6121	121.520	0.1112	4.71·10 ⁻⁴	5.6166	0.0796
45	2,2',4,4'	6.2910	8.1688	145.660	0.2853	3.00·10 ⁻⁴	6.0468	0.2442
46	2,2',4,5	5.7870	7.2785	136.050	0.0489	2.52·10 ⁻⁴	5.9821	0.1951
47	2,2',4,5'	6.2210	7.6295	141.160	0.1152	2.59·10 ⁻⁴	6.0646	0.1564
48	2,2',4,6	5.6370	6.6210	124.040	1.3559	4.02·10 ⁻⁴	5.8136	0.1766
49	2,2',4,6'	5.6370	6.7699	126.190	0.0853	3.62·10 ⁻⁴	5.7589	0.1219
50	2,2',5,5'	6.0910	6.6586	135.000	0.0994	2.62·10 ⁻⁴	6.1997	0.1087
51	2,2',5,6'	5.6270	6.1810	121.830	0.0035	9.77·10 ⁻⁶	5.8215	0.1945
52	2,2',6,6'	5.9040	5.3274	109.030	0.1105	4.85·10 ⁻⁴	5.6047	0.2993
53	2,3,3',4	6.1170	8.3206	146.050	0.0345	4.93·10 ⁻⁵	5.9921	0.1249
54	2,3,3',4'	6.1170	8.3589	146.460	0.0962	5.52·10 ⁻⁵	5.9982	0.1188
55	2,3,3',5	6.1770	7.8097	142.860	0.8606	7.89·10 ⁻⁵	6.1226	0.0544
56	2,3,3',5'	6.1770	7.9437	142.850	0.0479	5.54·10 ⁻⁵	6.0100	0.1670
57	2,3,3',6	5.9570	6.9758	129.260	0.0476	9.87·10 ⁻⁵	5.8151	0.1419
58	2,3,4,4'	5.4520	8.2526	145.460	0.0345	4.82·10 ⁻⁵	5.9947	0.5427
59	2,3,4,5	5.9430	6.7297	133.750	0.0286	3.72·10 ⁻⁵	6.1181	0.1751
60	2,3,4,6	5.8970	6.2938	123.830	0.0630	1.40·10 ⁻⁴	5.8617	0.0353
61	2,3,4',5	6.1770	8.0492	147.320	0.0969	6.32·10 ⁻⁵	6.1662	0.0108
62	2,3,4',6	5.9570	7.2049	133.110	0.0597	1.63·10 ⁻⁴	5.8873	0.0697
63	2,3,5,6	5.8670	5.6388	122.010	0.1451	9.11·10 ⁻⁵	6.0644	0.1974
64	2,3',4,4'	5.4520	8.6051	153.380	0.0353	5.40·10 ⁻⁵	6.1961	0.7441

ET36/2005 – Et. Finală/2006 – Lucrare in extenso

65	2,3',4,5	6.2070	8.1728	148.280	0.0359	5.11·10 ⁻⁵	6.1529	0.0541
66	2,3',4,5'	6.2670	8.1003	148.960	0.0334	5.87·10 ⁻⁵	6.2129	0.0541
67	2,3',4,6	6.0470	7.2353	133.270	0.0631	1.66·10 ⁻⁴	5.8817	0.1653
68	2,3',4',5	6.2310	7.5977	146.340	0.1246	5.56·10 ⁻⁵	6.3151	0.0841
69	2,3',4',6	5.9870	6.4091	128.760	0.0684	1.26·10 ⁻⁴	6.0319	0.0449
70	2,3',5,5'	6.2670	7.2087	143.450	0.0517	5.87·10 ⁻⁵	6.3459	0.0789
71	2,3',5',6	6.0470	6.0292	126.110	0.0503	1.18·10 ⁻⁴	6.0737	0.0267
72	2,4,4',5	6.6710	8.2735	152.290	0.0305	5.04·10 ⁻⁵	6.2873	0.3837
73	2,4,4',6	6.0570	7.4294	137.080	0.1203	4.40·10 ⁻⁴	5.9621	0.0949
74	2',3,4,5	6.1370	7.3073	139.090	0.0521	5.68·10 ⁻⁵	6.1119	0.0251
75	3,3',4,4'	6.5230	9.1490	161.430	0.2535	9.76·10 ⁻⁵	6.3366	0.1864
76	3,3',4,5	6.3570	8.1041	151.240	1.4919	1.77·10 ⁻⁴	6.4093	0.0523
77	3,3',4,5'	6.4270	8.5725	156.830	0.5807	1.28·10 ⁻⁴	6.3975	0.0295
78	3,3',5,5'	6.5830	7.9892	152.370	0.3196	2.56·10 ⁻⁴	6.4229	0.1601
79	3,4,4',5	6.3670	8.5262	157.620	0.1010	1.19·10 ⁻⁴	6.4188	0.0518
80	2,2',3,3',4	6.1420	8.2976	153.620	0.4084	3.31·10 ⁻⁴	6.3517	0.2097
81	2,2',3,3',5	6.2670	7.6890	148.730	0.7941	3.69·10 ⁻⁴	6.4172	0.1502
82	2,2',3,3',6	6.0410	6.9185	136.000	0.0442	3.88·10 ⁻⁴	6.1260	0.0850
83	2,2',3,4,4'	6.6110	8.4996	159.160	0.2075	3.32·10 ⁻⁴	6.4975	0.1135
84	2,2',3,4,5	6.2040	7.2394	145.350	0.1531	2.73·10 ⁻⁴	6.4160	0.2120
85	2,2',3,4,5'	6.3710	7.9248	154.060	0.0658	3.42·10 ⁻⁴	6.5038	0.1328
86	2,2',3,4,6	7.5160	6.7474	135.070	19.1930	3.56·10 ⁻⁴	7.4926	0.0234
87	2,2',3,4,6'	6.0770	6.8822	137.480	0.1178	8.98·10 ⁻⁴	6.1927	0.1157
88	2,2',3,4',5	6.3670	7.9310	154.600	0.1001	2.65·10 ⁻⁴	6.5303	0.1633
89	2,2',3,4',6	6.1370	7.1280	140.750	0.1762	3.36·10 ⁻⁴	6.2589	0.1219
90	2,2',3,5,5'	6.3570	7.3803	149.770	0.4452	3.11·10 ⁻⁴	6.5710	0.2140
91	2,2',3,5,6	6.0470	6.3200	132.670	0.5266	3.58·10 ⁻⁴	6.2654	0.2184
92	2,2',3,5,6'	6.1370	6.4532	134.510	0.0738	1.69·10 ⁻⁴	6.2662	0.1292
93	2,2',3,5',6	6.1370	6.6477	136.300	0.4964	7.70·10 ⁻⁴	6.2704	0.1334
94	2,2',3,6,6	5.7170	5.9325	123.370	0.0920	3.78·10 ⁻⁴	5.9865	0.2695
95	2,2',3',4,5	6.6710	7.8560	152.750	0.1031	3.34·10 ⁻⁴	6.4778	0.1932
96	2,2',3',4,6	6.1370	7.0705	139.710	0.1352	7.49·10 ⁻⁴	6.2187	0.0817
97	2,2',4,4',5	7.2110	8.1412	159.410	0.0577	2.85·10 ⁻⁴	6.6507	0.5603
98	2,2',4,4',6	6.2370	7.3053	144.740	0.2006	5.63·10 ⁻⁴	6.3537	0.1167
99	2,2',4,5,5'	7.0710	7.6259	154.880	0.1225	3.04·10 ⁻⁴	6.6712	0.3998
100	2,2',4,5,6'	6.1670	6.7261	138.380	0.0262	2.22·10 ⁻⁵	6.3246	0.1576
101	2,2',4,5',6	6.2270	6.8046	140.120	0.0279	5.27·10 ⁻⁵	6.3674	0.1404
102	2,2',4,6,6	5.8170	6.0829	126.540	0.0922	3.76·10 ⁻⁴	6.0634	0.2464
103	2,3,3',4,4'	6.6570	9.1546	168.730	0.0365	4.78·10 ⁻⁵	6.6435	0.0135
104	2,3,3',4,5	6.6470	8.2227	158.650	0.0384	3.80·10 ⁻⁵	6.5907	0.0563
105	2,3,3',4',5	6.7170	8.5882	164.350	0.0579	5.64·10 ⁻⁵	6.6895	0.0275
106	2,3,3',4,5'	6.7170	8.6053	163.610	0.0364	5.31·10 ⁻⁵	6.6482	0.0688
107	2,3,3',4,6	6.4870	7.5286	145.840	0.0862	1.67·10 ⁻⁴	6.3153	0.1717
108	2,3,3',4',6	6.5320	7.5906	148.130	0.3660	1.60·10 ⁻⁴	6.4101	0.1219
109	2,3,3',5,5'	6.7670	8.0899	159.630	0.0377	5.85·10 ⁻⁵	6.6891	0.0779
110	2,3,3',5,6	6.4570	7.1659	142.920	0.1454	9.27·10 ⁻⁵	6.3458	0.1112
111	2,3,3',5',6	6.5470	7.2338	144.320	0.0712	1.41·10 ⁻⁴	6.3721	0.1749
112	2,3,4,4',5	6.6570	8.2871	163.400	0.0322	3.75·10 ⁻⁵	6.7731	0.1161
113	2,3,4,4',6	6.4970	7.7007	150.180	0.1935	6.47·10 ⁻⁴	6.4241	0.0729
114	2,3,4,5,6	6.3040	6.1425	134.100	0.0371	1.31·10 ⁻⁴	6.3777	0.0737
115	2,3,4',5,6	6.4670	7.3123	146.830	0.3320	1.63·10 ⁻⁴	6.4673	0.0003
116	2,3',4,4',5	7.1210	8.7170	168.800	0.0387	4.74·10 ⁻⁵	6.8309	0.2901
117	2,3',4,4',6	6.5870	7.7307	152.260	0.0954	5.32·10 ⁻⁴	6.5009	0.0861

ET36/2005 – Et. Finală/2006 – Lucrare in extenso

118	2,3',4,5,5'	6.7970	8.1850	163.840	0.0386	$5.17 \cdot 10^{-5}$	6.8354	0.0384
119	2,3',4,5',6	6.6470	7.2928	148.060	0.1874	$2.93 \cdot 10^{-4}$	6.5149	0.1321
120	2',3,3',4,5	6.6470	8.3558	159.300	0.0339	$5.32 \cdot 10^{-5}$	6.5626	0.0844
121	2',3,4,4',5	6.7470	8.5776	165.470	0.0506	$1.30 \cdot 10^{-4}$	6.7401	0.0069
122	2',3,4,5,5'	6.7370	7.8399	159.710	0.0350	$5.59 \cdot 10^{-5}$	6.7977	0.0607
123	2',3,4,5,6'	6.5170	6.9321	142.900	0.2070	$7.11 \cdot 10^{-4}$	6.4244	0.0926
124	3,3',4,4'5	6.8970	9.2142	175.600	0.2341	$9.05 \cdot 10^{-5}$	6.9342	0.0372
125	3,3',4,5,5'	6.9570	8.6661	170.560	0.0774	$1.16 \cdot 10^{-4}$	6.9302	0.0268
126	2,2',3,3',4,4'	6.9610	9.3307	179.190	0.7685	$7.25 \cdot 10^{-4}$	7.0572	0.0962
127	2,2',3,3',4,5	7.3210	8.5439	169.150	0.0167	$1.44 \cdot 10^{-3}$	6.8653	0.4557
128	2,2',3,3',4,5'	7.3910	8.6910	172.480	0.0174	$1.04 \cdot 10^{-3}$	6.9659	0.4251
129	2,2',3,3',4,6	6.5870	7.6941	154.760	0.0299	$8.40 \cdot 10^{-4}$	6.6106	0.0236
130	2,2',3,3',4,6'	6.5870	7.8668	157.060	0.0471	$6.20 \cdot 10^{-4}$	6.6490	0.0620
131	2,2',3,3',5,5'	6.8670	7.8751	165.430	0.3205	$3.97 \cdot 10^{-4}$	7.0428	0.1758
132	2,2',3,3',5,6	7.3040	7.1724	150.190	0.2877	$1.26 \cdot 10^{-3}$	6.6304	0.6736
133	2,2',3,3',5,6'	7.1510	7.1833	151.550	0.0310	$1.92 \cdot 10^{-4}$	6.7081	0.4429
134	2,2',3,3',6,6'	6.5110	6.4063	138.560	0.0789	$2.98 \cdot 10^{-4}$	6.4604	0.0506
135	2,2',3,4,4',5'	7.4410	8.6591	174.490	0.3322	$6.17 \cdot 10^{-4}$	7.1058	0.3352
136	2,2',3,4,4',6	6.6770	7.7162	158.610	0.0244	$6.15 \cdot 10^{-4}$	6.7795	0.1025
137	2,2',3,4,4',6'	6.6770	7.8664	159.440	17.463	$4.04 \cdot 10^{-2}$	6.4781	0.1989
138	2,2',3,4,5,5'	7.5920	7.8362	166.540	0.1417	$4.20 \cdot 10^{-4}$	7.0949	0.4971
139	2,2',3,4,5,6	6.5170	6.7464	146.160	0.0415	$5.16 \cdot 10^{-4}$	6.6424	0.1254
140	2,2',3,4,5,6'	6.6070	6.9031	149.310	0.5610	$1.94 \cdot 10^{-4}$	6.7639	0.1569
141	2,2',3,4,5',6	6.6770	7.1854	153.300	0.1587	$1.13 \cdot 10^{-2}$	6.3768	0.3002
142	2,2',3,4,6,6'	6.2570	6.4146	138.950	0.0774	$2.91 \cdot 10^{-4}$	6.4743	0.2173
143	2,2',3,4',5,5'	6.8970	8.0823	169.690	0.0963	$6.14 \cdot 10^{-4}$	7.1201	0.2231
144	2,2',3,4',5,6	6.6470	7.3016	154.700	0.0982	$5.27 \cdot 10^{-4}$	6.7896	0.1426
145	2,2',3,4',5,6'	6.7370	7.4207	155.500	0.1185	$1.83 \cdot 10^{-4}$	6.7892	0.0522
146	2,2',3,4',5',6	7.2810	7.1955	153.590	0.0208	$1.48 \cdot 10^{-3}$	6.7440	0.5370
147	2,2',3,4',6,6'	6.3270	6.5980	141.640	0.0790	$3.00 \cdot 10^{-4}$	6.5158	0.1888
148	2,2',3,5,5',6	6.6470	6.7917	149.800	0.0450	$1.87 \cdot 10^{-4}$	6.7967	0.1497
149	2,2',3,5,6,6'	6.2270	6.0509	135.760	0.0815	$2.94 \cdot 10^{-4}$	6.4866	0.2596
150	2,2',4,4',5,5'	7.7510	8.1709	174.310	0.1986	$4.17 \cdot 10^{-4}$	7.3015	0.4495
151	2,2',4,4',5,6'	6.7670	7.4824	159.260	0.0188	$9.35 \cdot 10^{-5}$	6.9258	0.1588
152	2,2',4,4',6,6'	7.1230	6.7120	145.340	0.0787	$2.99 \cdot 10^{-4}$	6.6313	0.4917
153	2,3,3',4,4',5	7.1870	8.9450	180.720	0.0383	$2.82 \cdot 10^{-5}$	7.2624	0.0754
154	2,3,3',4,4',5'	7.1870	8.9295	180.060	0.0500	$3.70 \cdot 10^{-5}$	7.2402	0.0532
155	2,3,3',4,4',6	7.0270	8.1071	166.380	0.3064	$2.87 \cdot 10^{-4}$	6.9903	0.0367
156	2,3,3',4,5,5'	7.2470	8.4424	175.580	0.0401	$3.17 \cdot 10^{-5}$	7.2467	0.0003
157	2,3,3',4,5,6	6.9370	7.3571	156.700	0.0739	$1.34 \cdot 10^{-4}$	6.8677	0.0693
158	2,3,3',4,5',6	7.0870	7.6904	161.910	1.6216	$2.19 \cdot 10^{-4}$	7.0623	0.0247
159	2,3,3',4',5,5'	7.2470	8.1738	174.530	0.0248	$3.74 \cdot 10^{-5}$	7.3121	0.0651
160	2,3,3',4',5,6	6.9970	7.5592	161.980	0.2656	$9.21 \cdot 10^{-5}$	7.0309	0.0339
161	2,3,3',4',5',6	7.0270	7.2468	158.640	0.1114	$1.16 \cdot 10^{-4}$	7.0031	0.0239
162	2,3,3',5,5',6	7.0570	7.2209	157.930	0.1685	$9.63 \cdot 10^{-5}$	6.9874	0.0696
163	2,3,4,4',5,6	6.9370	7.4614	161.140	0.5263	$4.59 \cdot 10^{-4}$	7.0393	0.1023
164	2,3',4,4',5,5'	7.2770	8.4739	180.990	0.0537	$4.30 \cdot 10^{-5}$	7.4731	0.1961
165	2,3',4,4',5',6	7.1170	7.4940	163.550	0.2913	$5.09 \cdot 10^{-4}$	7.1138	0.0032
166	3,3',4,4',5,5'	7.4270	9.1962	189.460	0.0633	$8.19 \cdot 10^{-5}$	7.5426	0.1156
167	2,2',3,3',4,4',5	7.2770	8.9382	188.200	0.1129	$9.03 \cdot 10^{-5}$	7.5986	0.3216
168	2,2',3,3',4,4',6	6.7040	8.1822	173.820	0.1327	$1.78 \cdot 10^{-3}$	7.2194	0.5154
169	2,2',3,3',4,5,5'	7.3370	8.4591	183.610	0.0758	$9.65 \cdot 10^{-4}$	7.5620	0.2250
170	2,2',3,3',4,5,6	7.0270	7.6027	165.040	0.0482	$9.50 \cdot 10^{-4}$	7.1005	0.0735

ET36/2005 – Et. Finală/2006 – Lucrare in extenso

171	2,2',3,3',4,5,6'	7.1170	7.5973	167.060	0.0124	8.87·10 ⁻⁵	7.2218	0.1048
172	2,2',3,3',4,5',6	7.1770	7.7165	169.220	0.0266	8.26·10 ⁻⁵	7.2683	0.0913
173	2,2',3,3',4,6,6'	6.7670	7.0135	155.320	0.0680	2.45·10 ⁻⁴	6.9467	0.1797
174	2,2',3,3',4',5,6	7.0870	7.6992	169.170	0.0578	2.24·10 ⁻⁴	7.2703	0.1833
175	2,2',3,3',5,5',6	7.1470	7.3049	165.140	0.1383	2.63·10 ⁻⁴	7.2622	0.1152
176	2,2',3,3',5,6,6'	6.7370	6.6644	151.780	0.0715	2.45·10 ⁻⁴	6.9374	0.2004
177	2,2',3,4,4',5,5'	7.3670	8.6195	188.140	0.0550	4.81·10 ⁻³	7.5491	0.1821
178	2,2',3,4,4',5,6	7.1170	7.7127	170.270	6.8760	6.77·10 ⁻³	7.5427	0.4257
179	2,2',3,4,4',5,6'	7.2070	7.8062	171.680	0.6290	2.78·10 ⁻⁴	7.3739	0.1669
180	2,2',3,4,4',5',6	7.2070	7.8790	173.190	0.1139	3.01·10 ⁻⁴	7.3733	0.1663
181	2,2',3,4,4',6,6'	6.8570	7.1523	158.990	0.0677	2.43·10 ⁻⁴	7.0505	0.1935
182	2,2',3,4,5,5',6	7.9330	6.8634	164.000	0.4577	1.39·10 ⁻⁵	7.4292	0.5038
183	2,2',3,4,5,6,6'	6.6970	6.2008	148.390	0.0693	2.38·10 ⁻⁴	6.9828	0.2858
184	2,2',3,4',5,5',6	7.1770	7.3776	168.560	0.0448	1.14·10 ⁻⁴	7.3819	0.2049
185	2,2',3,4',5,6,6'	6.8270	6.7527	155.210	0.0714	2.43·10 ⁻⁴	7.0519	0.2249
186	2,3,3',4,4',5,5'	7.7170	9.1221	195.930	0.0488	3.59·10 ⁻⁵	7.8604	0.1434
187	2,3,3',4,4',5,6	7.4670	8.2487	178.880	0.5202	6.32·10 ⁻⁴	7.4850	0.0180
188	2,3,3',4,4',5',6	7.5570	8.2502	179.590	0.6986	6.92·10 ⁻⁴	7.5259	0.0311
189	2,3,3',4,5,5',6	7.5270	7.7471	173.540	0.0626	2.84·10 ⁻⁴	7.4413	0.0857
190	2,3,3',4',5,5',6	7.5270	7.6407	174.490	0.4337	1.22·10 ⁻⁴	7.5600	0.0330
191	2,2',3,3',4,4',5,5'	8.6830	8.8559	201.230	0.0917	6.13·10 ⁻⁴	8.1879	0.4951
192	2,2',3,3',4,4',5,6	7.5670	8.1340	185.280	0.0365	5.66·10 ⁻⁵	7.8039	0.2369
193	2,2',3,3',4,4',5',6	7.6570	8.0500	185.640	0.0407	3.72·10 ⁻⁵	7.8562	0.1992
194	2,2',3,3',4,4',6,6'	7.3070	7.5022	173.290	0.0595	1.97·10 ⁻⁴	7.5363	0.2293
195	2,2',3,3',4,5,5',6	7.6270	7.7337	180.830	0.0627	1.47·10 ⁻⁴	7.7742	0.1472
196	2,2',3,3',4,5,6,6'	7.2070	6.9867	166.280	0.0619	1.96·10 ⁻⁴	7.4437	0.2367
197	2,2',3,3',4,5',6,6'	7.2770	7.1190	169.460	0.0626	1.98·10 ⁻⁴	7.5285	0.2515
198	2,2',3,3',4',5,5',6	7.6270	7.6994	180.700	0.0627	1.52·10 ⁻⁴	7.7827	0.1557
199	2,2',3,3',5,5',6,6'	8.4230	6.6004	165.330	0.0656	1.96·10 ⁻⁴	7.5645	0.8585
200	2,2',3,4,4',5,5',6	7.6570	7.7846	184.510	0.6727	3.23·10 ⁻⁴	7.9513	0.2943
201	2,2',3,4,4',5,6,6'	7.3070	7.0818	170.100	0.0617	1.93·10 ⁻⁴	7.5726	0.2656
202	2,3,3',4,4',5,5',6	8.0070	8.1803	191.360	1.1642	5.30·10 ⁻⁴	8.1139	0.1069
203	2,2',3,3',4,4',5,5',6	9.1430	7.9885	197.410	0.0270	6.07·10 ⁻⁵	8.4003	0.7427
204	2,2',3,3',4,4',5,6,6'	7.7470	7.4690	184.980	0.0550	1.57·10 ⁻⁴	8.0680	0.3210
205	2,2',3,3',4,5,5',6,6'	8.1640	7.1318	180.950	0.0577	1.60·10 ⁻⁴	8.0319	0.1321
206	2,2',3,3',4,4',5,5',6,6'	9.6030	7.4035	197.030	0.0512	1.33·10 ⁻⁴	8.6287	0.9743

Three molecular descriptors take into consideration the geometry of PCBs (IIDDKGg, IHDRKEg, and aSMMjQg) and one the topology of compounds (aHMmjQt). As atomic property, two descriptors consider the partial charge (aHMmjQt, and aSMMjQg), one the group electronegativity (IIDDKGg) and one the atomic electronegativity (IHDRKEg). Looking at the interaction descriptor (the fifth letter in descriptors name) it can be observed that all descriptors consider the elastic force.

The results of multiple linear regressions associated to the four-varied model (see table 2, and table 3) sustain the estimation and prediction abilities of the best performing SAR model.

ET36/2005 – Et. Finală/2006 – Lucrare in extenso

In the table 3 are the 95% probability of confidence intervals - lower ($95\%CI_L$) and upper ($95\%CI_U$) boundaries, coefficients, standard error (StdErr) of the coefficient, Student test parameter (t) and Student probability (p_t).

Table 2. Statistics associated with the tetra-varied model

Characteristic	Notation	Values
Correlation coefficient	r	0.9575
Squared correlation coefficient	r^2	0.9168
Adjusted squared correlation coefficient	r^2_{adj}	0.9151
Standard error of estimated	S_{est}	0.2420
Fisher parameter	F_{est}	554
Probability of wrong model	$p_{est}(\%)$	$< 1 \cdot 10^{-15}$
Cross-validation leave-one-out (loo) score	$r^2_{cv(loo)}$	0.9093
Fisher parameter for loo analysis	F_{pred}	504
Probability of wrong model for loo analysis	$p_{pred}(\%)$	$< 1 \cdot 10^{-15}$
Standard error for loo analysis	S_{loo}	0.2526
The difference between r^2 and $r^2_{cv(loo)}$	$r^2 - r^2_{cv(loo)}$	0.0075
Squared correlation coefficients between each descriptor and measured octanol-water partition coefficients or between pairs of descriptors	$r^2(IIDDKGg, IHDRKEg)$	0.48245
	$r^2(IIDDKGg, aHMmjQt)$	0.00005
	$r^2(IIDDKGg, aSMMjQg)$	0.00385
	$r^2(IHDRKEg, aHMmjQt)$	0.00039
	$r^2(IHDRKEg, aSMMjQg)$	0.00073
	$r^2(aHMmjQt, aSMMjQg)$	0.24805
	$r^2(IIDDKGg, \log_{Kow})$	0.15111
	$r^2(IHDRKEg, \log_{Kow})$	0.78907
	$r^2(aSMMjQg, \log_{Kow})$	0.00932
	$r^2(aHMmjQt, \log_{Kow})$	0.00786

Table 3. Statistics associated with the four-varied model

	$95\%CI_L$	Coefficients	$95\%CI_U$	StdError	t	p_t (%)
Intercept	2.735	3.039	3.343	0.154	19.716	$7.27 \cdot 10^{-47}$
IIDDKGg	-0.477	-0.421	-0.365	0.028	-14.804	$5.09 \cdot 10^{-32}$
IHDRKEg	0.042	0.044	0.046	0.001	41.725	$5.97 \cdot 10^{-99}$
aHMmjQt	0.049	0.070	0.090	0.010	6.639	$2.89 \cdot 10^{-8}$
aSMMjQg	-47.601	-37.499	-27.397	5.123	-7.319	$5.86 \cdot 10^{-10}$

The model which consider in estimation four molecular descriptors is significant statistically, having a probability of a wrong model less than $1 \cdot 10^{-15}$ (%). The estimation ability of the SAR model is sustained by the value of the correlation coefficient ($r = 0.9575$), the confidence boundaries associated with the coefficients (see table 3), and probabilities associated with Student tests (for all coefficients less than 0.001 - see table 3). Almost ninety-two percent ($r^2 = 0.9168$) from variation of octanol-water partition coefficient can be explained by its linear relationship with the variation of the four molecular descriptors used in the model. The probability of wrong model for leave-one-out analysis ($p_{pred}(\%) < 1 \cdot 10^{-15}$) and

ET36/2005 – Et. Finală/2006 – Lucrare in extenso

its associated Fisher parameter ($F_{\text{pred}} = 504$) sustains the estimation ability of the model. The four-varied SAR model is a stable one, stability sustained by the values of difference between correlation coefficient and cross validation leave-one-out correlation score ($r^2 - r_{\text{cv}(100)}^2 = 0.0075$). The power of the four-varied model in octanol-water partition coefficient prediction of PCBs is sustained by the absence of multicollinearity of descriptors used by the model (see the squared correlation coefficients between pairs of descriptors, which always is less than 0.48 - table 2).

The plot of dependency between measured ($\log K_{\text{ow}}$) and estimated based on the structure of polychlorinated biphenyls compounds obtained with the tetra-varied model is in figure 1.

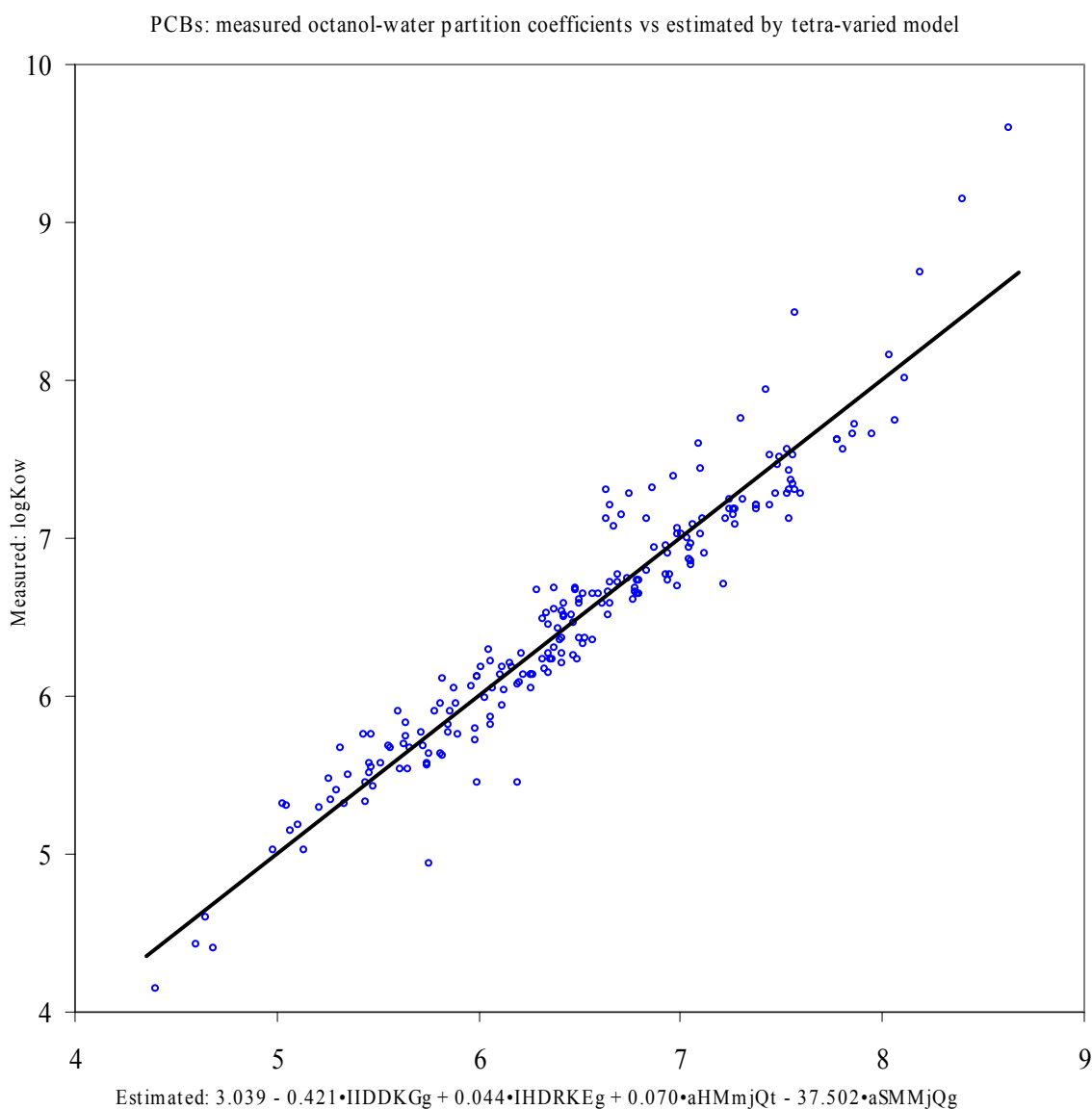


Figure 1. Measured vs. estimated $\log K_{\text{ow}}$ by the tetra-varied model

ET36/2005 – Et. Finală/2006 – Lucrare in extenso

The estimation values of octanol-water partition coefficients by the use of the four-varied model of are less or greater than measured values (see figure 2). Note that, the mean and 95% confidence intervals of the mean and standard error for measured ($m_{\text{Measured}} = 6.4802$, $95\%CI_{\text{Measured}} = [6.3709, 6.5895]$, $StdErr_{\text{Measured}} = 0.0554$) and estimated ($m_{\text{Estimated}} = 6.4806$, $95\%CI_{\text{Estimated}} = [6.3664, 6.5947]$, $StdErr_{\text{Estimated}} = 0.0579$) octanol-water partition coefficients are almost equal.

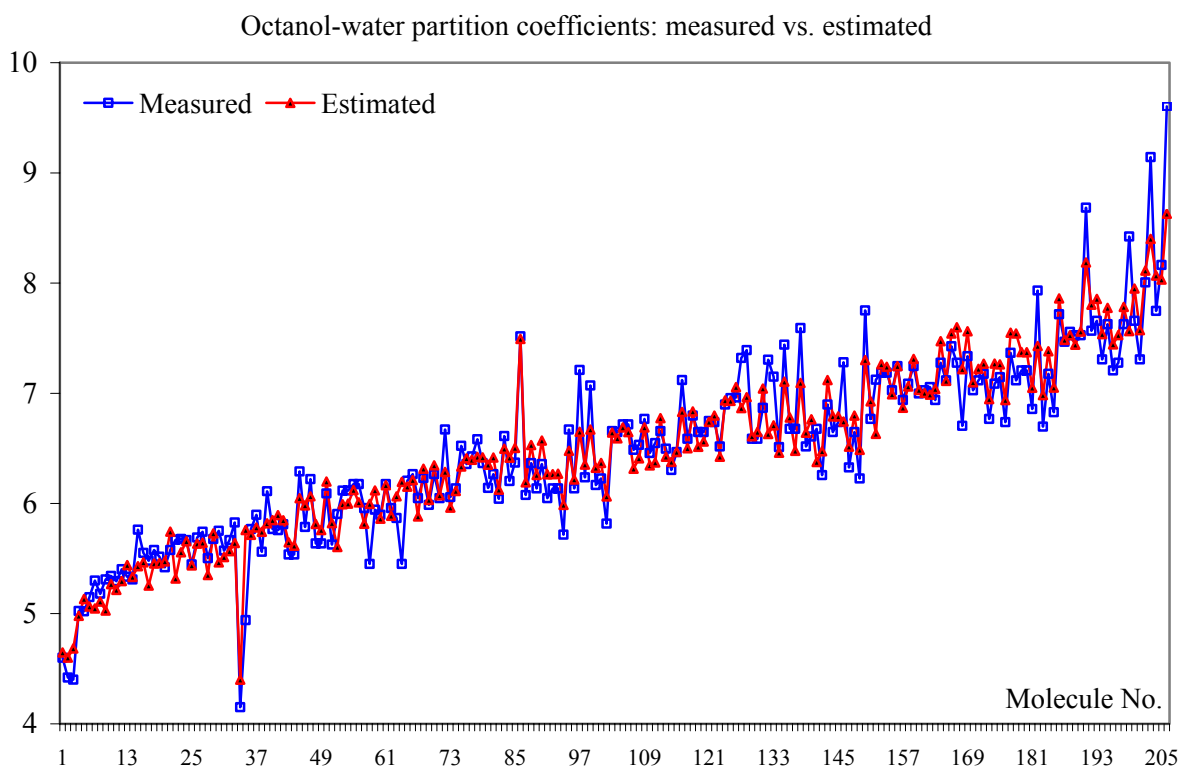


Figure 2. Variation of measured (blue line) and estimated (red line) by the four-varied model of octanol-water partition coefficient for PCBs

In order to see the estimation abilities of four-varied model, measured and estimated values were sort by the absolute differences between estimated and measured octanol-water partition coefficient of PCBs and split into two subsets (first containing one-hundred PCBs and second containing the other one-hundred and six PBCs). The graphical representations are in figure 3a (one-hundred compounds) and 3b (one-hundred and six compounds), where the PCB number was associated with corresponding estimated and measured values.

All squared correlation coefficients in training as well as in test sets are greater than 0.9, sustaining the prediction ability of the four-varied model. More, the mean of squared correlation coefficients in test sets is a little bit higher compared with the mean of squared correlation coefficient in training sets, and the dispersions of squared correlation coefficients

ET36/2005 – Et. Finală/2006 – Lucrare in extenso

are very small for both sets. All the regressions in training and test sets are highly significant ($p < 0.001$).

Analyzing the regressions coefficients it can be observed that with no exception the values of coefficients respect the 95% confidence intervals associated to the four-varied model (see table 3 and table 4). More, as it is expected, the 95% CI values (table 4) obtained in training and test sets analyses are contained by the 95% CI values of four-varied model (table 3).

The plot of measured vs. estimated octanol-water partition coefficients in training set (blue line and dots) of sample size equal one-hundred thirty-seven (corresponding with 2/3 from total sample of PCBs) and corresponding test set (red line and dots) of sample size equal with sixty-nine (1/3 from total sample of PCBs) is in figure 5.

Analyzing the residuals of the four-varied model allowed us to assess the suitability of the model. Looking at the differences between measured and estimated octanol-water partition coefficient for PCBs (figure 4) it can be observed that the values vary around zero and most of them between -0.5 and 0.5.

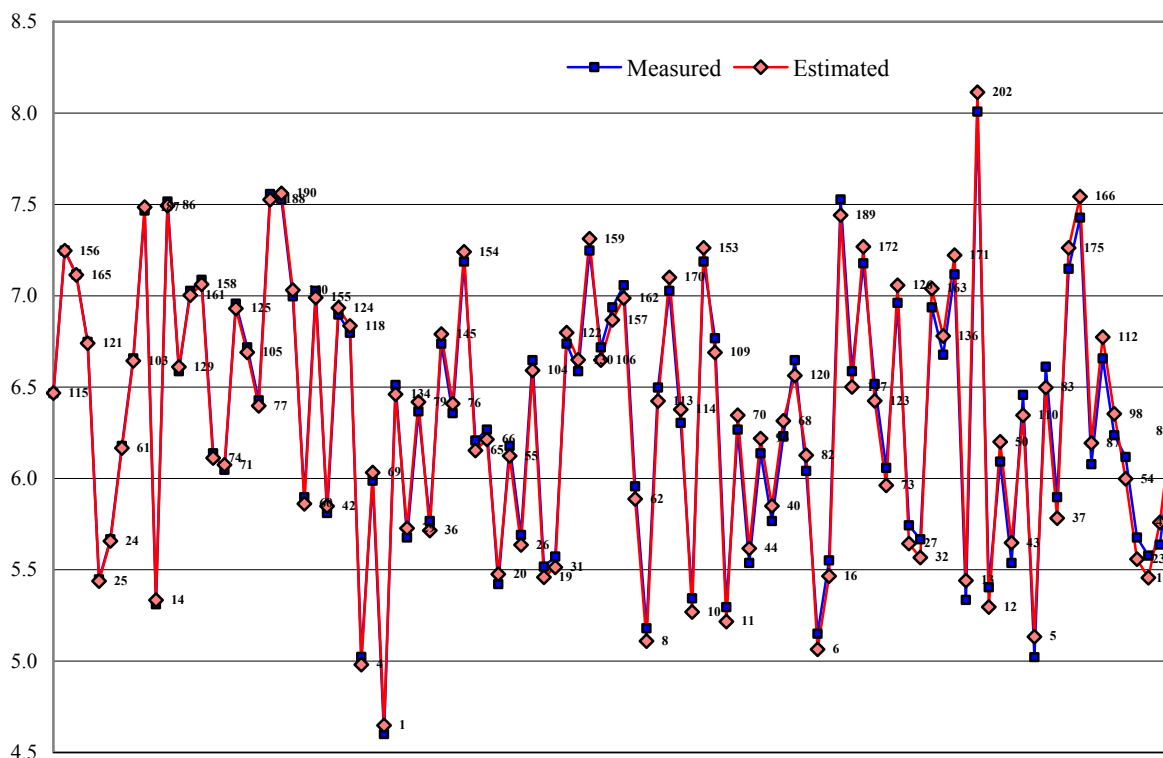


Figure 3a. Measured (blue line) and estimated by the tetra-varied model (red-line) of octanol-water partition coefficients for one-hundred PCBs

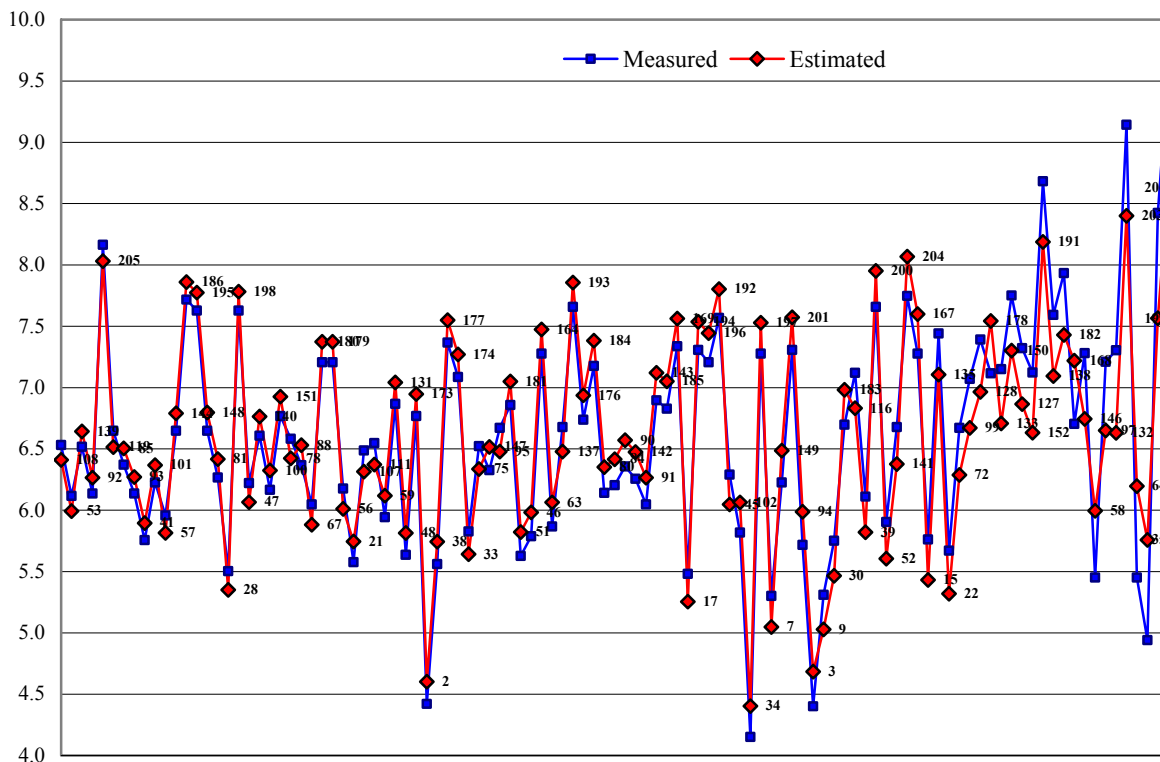


Figure 3b. Measured (blue line) and estimated by the tetra-varied model (red line) of octanol-water partition coefficients for one-hundred and six PCB

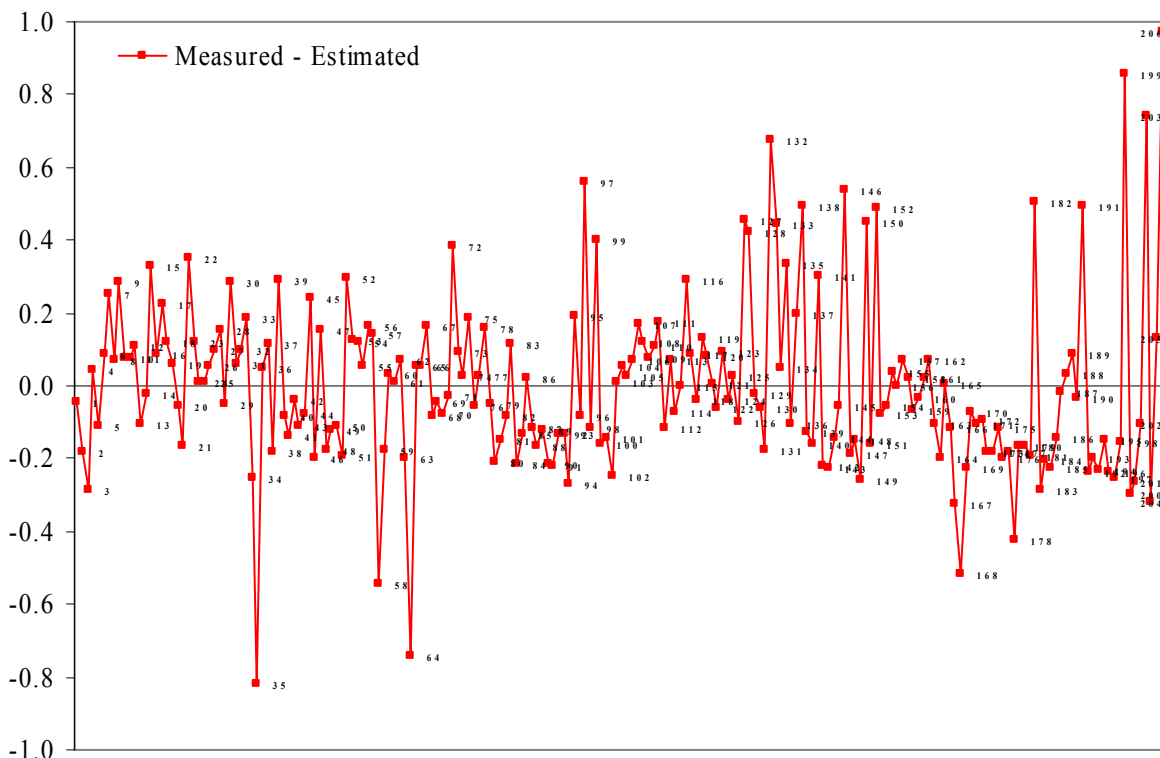


Figure 4. The differences between measured and estimated by the tetra-varied model of octanol-water partition coefficients for PCBs

The prediction abilities of the four-varied SAR model were studied through training and test sets analysis, and the results are in table 4. There were analyzed twelve situations,

ET36/2005 – Et. Finală/2006 – Lucrare in extenso

starting with a training sample size equal with 116 and increasing the number of PCBs included into training sets through randomization with seven until one hundred ninety-three. In table 4, there were included the number of PCBs in training sets (No_{tr}), the coefficients of the model, the squared correlation coefficient for training set (r_{tr}^2), Fisher parameter associated with training set regression (F_{tr}), the number of the PCBs in test sets (No_{ts}), the squared correlation coefficient for test set (r_{ts}^2), Fisher parameter associated with training set regression (F_{ts}), the mean (Mean) and standard deviation (StDev) for squared correlation coefficients and the 95% probability CI [$95\%CI_L$ and $95\%CI_U$] for coefficients.

Table 4. Results of training vs. test sets analysis

No_{tr}	intercept	IIDDKGg	IHDRKEg	aHMmjQt	aSMMjQg	r_{tr}^2	F_{tr}	No_{ts}	r_{ts}^2	F_{ts}
116	3.070	-0.408	0.043	0.064	-34.937	0.9141	295*	90	0.9219	235*
123	3.058	-0.390	0.043	0.064	-43.454	0.9229	353*	83	0.9043	176*
130	2.957	-0.413	0.044	0.067	-33.462	0.9232	376*	76	0.9068	169*
137	3.011	-0.438	0.045	0.064	-32.008	0.9004	298*	69	0.9432	256*
144	3.090	-0.450	0.045	0.062	-45.236	0.9143	371*	62	0.9186	148*
151	3.102	-0.432	0.044	0.062	-42.983	0.9173	405*	55	0.9075	122*
158	3.137	-0.460	0.046	0.073	-37.319	0.9200	440*	48	0.9041	82*
165	3.091	-0.428	0.044	0.070	-37.661	0.9143	427*	41	0.9247	110*
172	3.063	-0.426	0.044	0.069	-36.945	0.9161	456*	34	0.9202	83*
179	3.085	-0.429	0.044	0.069	-37.219	0.9098	439*	27	0.9582	106*
186	2.990	-0.420	0.045	0.070	-37.650	0.9090	452*	20	0.9876	178*
193	3.067	-0.430	0.044	0.074	-37.466	0.9160	513*	13	0.9249	24*

* $p < 0.001$

$95\%CI_L$	3.028	-0.439	0.044	0.065	-40.566	0.9148	Mean	0.9268
$95\%CI_U$	3.092	-0.415	0.045	0.070	-35.490	0.0063	StDev	0.0250

Starting with the above describe model, and by the use of the original software [61], the octanol-water partition coefficient of new polychlorinated biphenyls can be obtains in a short time, without any experiments, following the next steps: drawing by the use of HyperChem software the three dimensional structure of the new PCB, choosing the model of prediction from the list (in our case PCB_lkow), browsing the *.hin file, and computing the octanol-water partition coefficient based on the four-varied SAR equation.

[61] ***, *MDF SAR Predictor*, © 2005, Virtual Library of Free Software, available at: http://vl.academicdirect.org/molecular_topology/mdf_findings/sar.

Training (137 PCBs) vs test (69 PCBs) analysis:
measured vs estimated by four-varied model of octanol-water partition coefficient

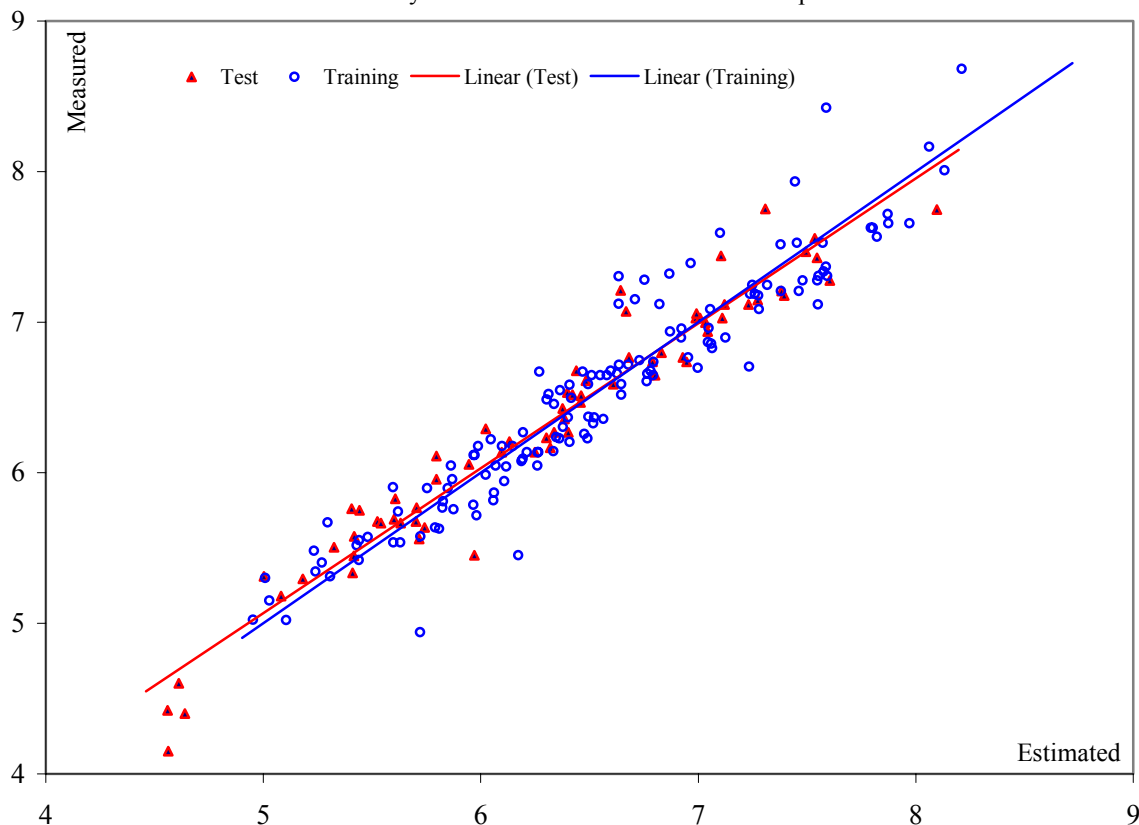


Figure 4. Training (137 PCBs) vs test (69 PCBs) analysis with four-varied model

Conclusions

Modeling the octanol-water partition coefficient of polychlorinated biphenyls by integration of complex structural information provide a stable and performing four-varied model, allowing us to make remarks about relationship between structure of PCBs and associated octanol-water partition coefficients. Thus, the octanol-water partition coefficient of studied PCBs is like to be of geometry and topology nature, depending by the partial change, group and atomic electronegativity as atomic properties, and being in relation with the elastic force.

Concluzii

Molecular Descriptors Family (MDF), așa cum a fost construită, își dovedește abilități corelaționale deosebite.

Practic pentru toate seturile investigate, rezultatele au fost superioare celor raportate anterior în literatura de specialitate, folosind alte metode de corelare a structurii cu activitatea biologică.

Rezultatele obținute sunt disponibile online la adresa:

http://vl.academicdirect.org/molecular_topology/

Cluj-Napoca,

la 28.10.2006

Director temă ET36/2005,

Șef. L. Dr., Ing. Lorentz JÄNTSCHI

<http://lori.academicdirect.org>
